Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Big Data as a Data Source for Official Statistics: experiences at Statistics Netherlands

Piet J.H. Daas, Marco Puts, Martijn Tennekes, and Alex Priem[1]

Abstract

More and more data are being produced by an increasing number of electronic devices physically surrounding us and on the internet. The large amount of data and the high frequency at which they are produced have resulted in the introduction of the term 'Big Data'. Because of the fact that these data reflect many different aspects of our daily lives and because of their abundance and availability, Big Data sources are very interesting from an official statistics point of view. However, first experiences obtained with analyses of large amounts of Dutch traffic loop detection records, call detail records of mobile phones and Dutch social media messages reveal that a number of challenges need to be addressed to enable the application of these data sources for official statistics. These and the lessons learned during these initial studies will be addressed and illustrated by examples. More specifically, the following topics are discussed: the three general types of Big Data discerned, the need to access and analyse large amounts of data, how we deal with noisy data and look at selectivity (and our own bias towards this topic), how to go beyond correlation, how we found people with the right skills and mind-set to perform the work, and how we have dealt with privacy and security issues.

Key Words: Big Data, Official statistics, Challenges, Lessons learned.

1. Introduction

In our modern digital era, data nearly touches every aspect of our lives, from the way we shop on the web, travel by car or public transport, search product information and communicate with friends and family. In addition to this, our roundabouts are captured by cameras, mobile phones and wireless local area networks. All these data are stored and can potentially be harvested. However, in their raw form these Big Data sources are not immediately valuable. One must be able to separate the signal from the noise, i.e. have statistical expertise in deriving information from large amounts of data, to extract their meaning. Here, knowledge in statistical inference from Big Data is needed (London Workshop, 2014). This is a relatively new area of expertise: the area of valid statistical analysis for Big Data is only just emerging (Fan et al., 2014). The challenge of these kind of analyses is to extract the signal (if present) relevant for the topic of interest from a large and (very) noisy data set (Silver, 2010).

Big Data is a very interesting source for official statistics (Glasson et al., 2013) as it enables the potential production of speedy and considerable relevant official figures at relatively low costs. How this can be achieved in practice is a topic of interests for many National Statistical Institutes. A number of challenges have been identified (more in section 2). For instance, many Big Data sources are composed of observational data and, as a consequence, have no well-defined target population, often lack structure and are of varying quality. This makes it difficult to apply traditional statistical methods, based on sampling theory. However, not every Big Data source faces the same issues. By studying a number of Big Data sources, i.e. road sensor data, call detail records of mobile phones and social media messages, the group of Big Data researchers at Statistics Netherlands are obtaining insight into the study of these sources, learn what works and doesn't work and get valuable insight into the potential application of Big Data for official statistics. This paper provides an overview of these findings.

[1]All authors are employees of Statistics Netherlands. Contact person: Piet Daas, CBS-weg 11 Heerlen, the Netherlands, 6412 EX (pjh.daas@cbs.nl). The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

2. Challenges

A number of challenges have been identified that need to be addressed when starting to use Big Data for official statistics (Daas and Van der Loo, 2013; Glasson et al., 2013; Struijs et al., 2014). Below an overview is given of the main ones.

2.1 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 data processors, who have their own concerns regarding costs, confidentiality and other issues. However, they might also benefit from cooperating with statistical agencies, for instance by way of the quality feedback NSI's provide. Terms and conditions have to be negotiated that are acceptable to both official statisticians and data providers.

2.2 Privacy

Privacy protection of individuals is 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.

2.3 Methodology

Many Big Data sources are composed of event-driven observational data which are not designed for traditional statistical analysis. They lack well-defined target populations, data structures and quality guarantees. This makes it hard to apply statistical methods based on sampling theory (Daas and Puts, 2014a). For example, assessing selectivity issues is challenging (Buelens et al., 2014). Since an increasing number of Big Data sources are text-based or composed of images, the need to extract information from these kinds of 'data' sources increases. This calls for information extraction methods, such as text mining and machine learning techniques, not yet very familiar to official statisticians; although they have already been identified several years ago (Fyhrlund et al., 2005; Saporta, 2000).

2.4 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 (Nguyen et al., 2013); but a lot still needs to be done (Daas and Burger, 2014).

2.5 Technology

An obvious challenge is the processing, storage and transfer of large data sets. Technological advances in the area of High Performance Computing 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 (Hager and Wellein, 2010). The technological challenges include security mechanisms, which makes for example cheap cloud-based solutions not an option for NSIs.

2.6 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. This poses a risk for continuity and a need for a more flexible way of working.

3. Big Data studies

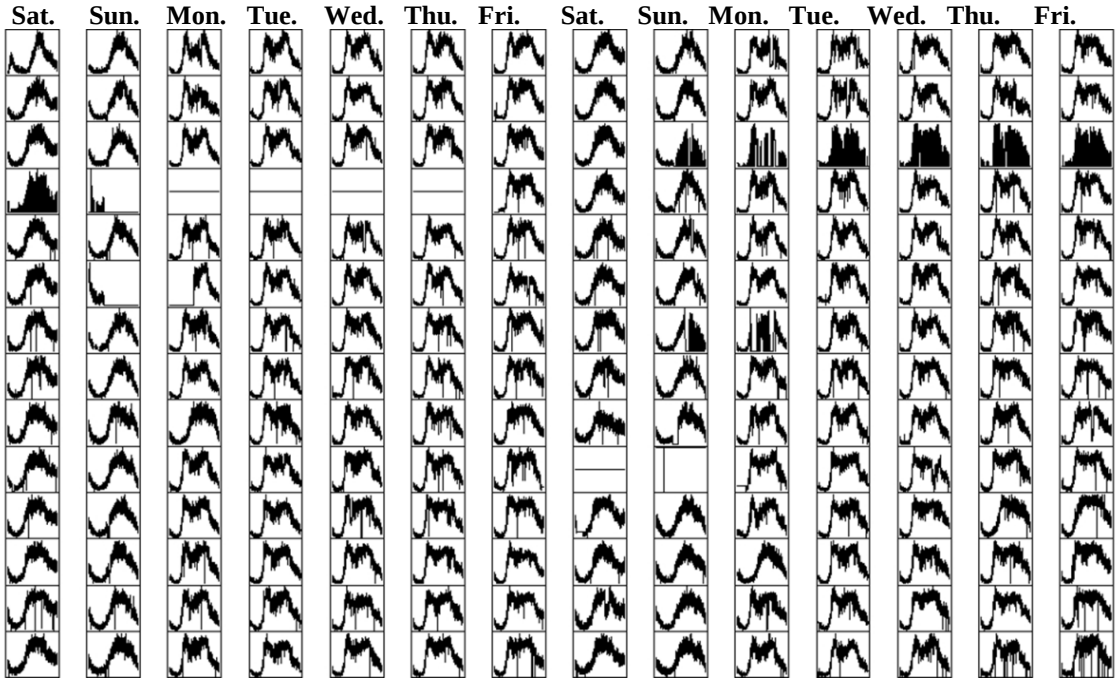
3.1 Sources

In this chapter, we discuss three typical examples of Big Data research conducted at Statistics Netherlands. Other Big Data related studies performed at our office include internet robots, scanner data and satellite images. Further opportunities like analysing financial transactions 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 has been 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 the larger administrative sources like the population register, VAT data and wages and salaries records could be interpreted as such. Looking at these more traditional sources from a Big Data point of view may provide new insights.

3.2 Road sensor data

In the Netherlands, there are more than 60,000 road sensors of which 20,000 are positioned on the Dutch highways. These sensors detect the number of passing vehicles in various length classes each minute. This results in a total of 230 million records a day for the Highway sensors alone. 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 to Statistics Netherlands. Since the data cannot be related back to individual vehicles, privacy concerns do not apply. The latter makes this data set attractive for experimentation. The most important issue we ran into while studying road sensor data was the fact that the quality of the data fluctuates tremendously. For some sensors, data for many minutes are not available and, because of the stochastic nature of the arrival times of vehicles at a road sensor, it is hard to directly derive the number of vehicles missing during these minutes. For this purpose, an adaptive filter was developed that is tuned to the stochastic behaviour of the arrival times of the vehicles at the sensor (Puts et al., 2014). The quality of the data not only varies per minute but also per day (Fig. 3.2-1). By correcting for missing data and combining the daily profiles provided by sensors on the same road sections, the coverage and quality of the data is improved. In this way we are able to make traffic indices that describe the regional situation, at the NUTS-3 level, on the Dutch roads. Combining these regional findings gives a very good impression of the state of the country concerning road traffic (Daas et al., 2014).

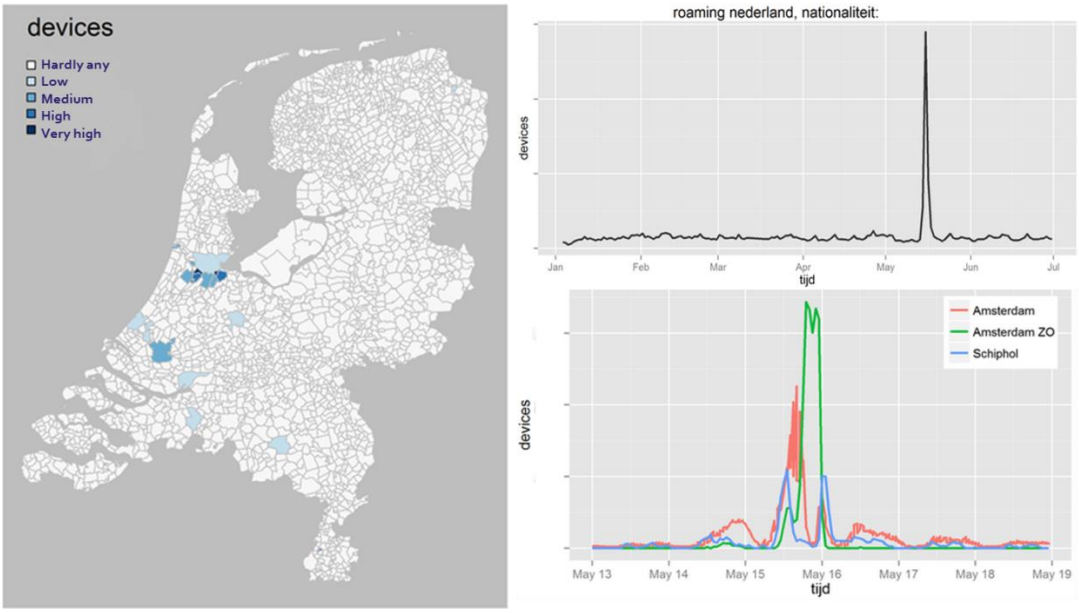
Figure 3.2-1
Daily profiles of a road sensor on the IJsselmeer dam ("Afsluitdijk") during 196 subsequent days.



3.3 Mobile phone data

Nowadays, people carry mobile phones with them everywhere and use their phones often throughout the day. To manage the phone traffic, a lot of data needs to be processed by mobile phone companies. These data are 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. Through a three-party contract, Statistics Netherlands got access to call detail records (CDR) data from a Dutch mobile phone company with a market share of approximately one third of the Dutch mobile phone market. The CDR data amounts to 115 million records a day and contains information on both Dutch and roaming users of their network. The anonymized CDR micro data were processed by a specialized intermediate company, according to queries specified by Statistics Netherlands. Only aggregated results were forwarded to Statistics Netherlands as agreed to protect privacy. Several uses for official statistics were studied, including inbound tourism (Heerschap et al., 2014) and daytime population (Tennekes and Offermans, 2014). In Figure 3.3-1 an example is shown of CDR-data applied to inbound 'tourism'. In this figure the activity of mobile phones assigned to one of the European countries involved in the European League Final, held in 2013 on May 15th, in the Amsterdam arena. The most striking finding is the fact that the activity of mobile phones from that particular country around that date is much higher than the activity of such phones in the remainder of the period studied. It nicely illustrates tourists visiting our country for a particular event during a very short period (Heerschap et al., 2014). It is highly likely that the majority of these visitors are not included in the official, accommodation based, tourism statistics.

Figure 3.3-1
Mobile phone activity of phones registered in a single European country around the period of the UEFA Europa League Final in the Amsterdam Arena in 2013. Relative activity in all and specific Dutch regions are shown, including the overall activity in the Netherlands during the first half of the year.



In Figure 3.3-2 an example of the daytime population studies is shown. 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. In the figure, the daytime population at noon on a May Monday for the five largest Dutch municipalities is shown. Findings for a typical working municipality (Haarlemmermeer, where Schiphol Airport is located) and a typical commuter municipality ('Almere') are also included. Colours indicate the number of people that leave, stay or enter. Dots indicate the officially registered population. This work is an example of a potential new Big Data-based statistic that an NSI could produce (Tennekes and Offermans, 2014).

3.4 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 (Daas et al., 2012). The latter also negatively affected text mining studies. Coosto-data studies revealed that the sentiment in Dutch social media messages was highly correlated with Dutch consumer confidence (Fig. 3.4-1). This phenomenon is predominantly affected by changes in the sentiment of all Dutch public Facebook messages (r = 0.85). The inclusion of various selections of public Twitter messages improved this association and the response to changes in sentiment (up to r = 0.89). The observed sentiment was stable on a monthly and weekly basis, but daily figures displayed highly volatile behaviour.

Figure 3.3-2
Estimated daytime population at noon on a May Monday for the five largest Dutch municipalities.

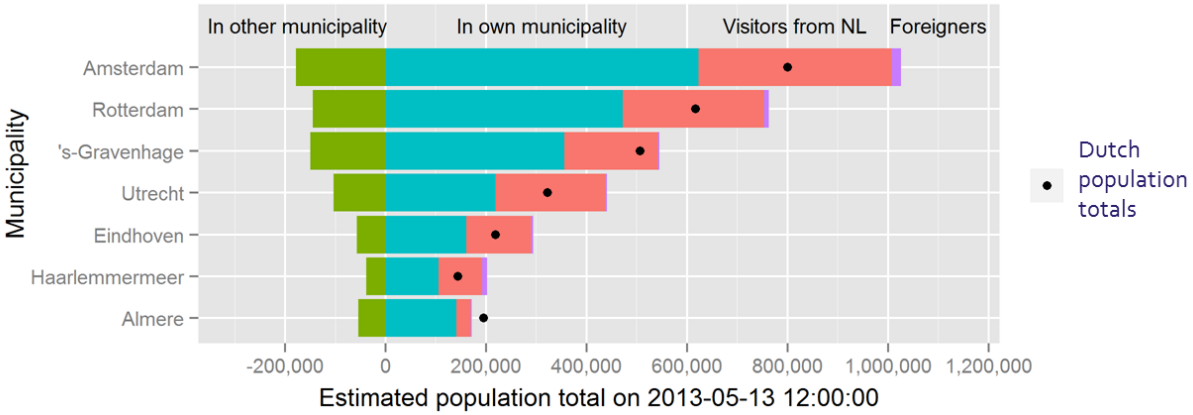
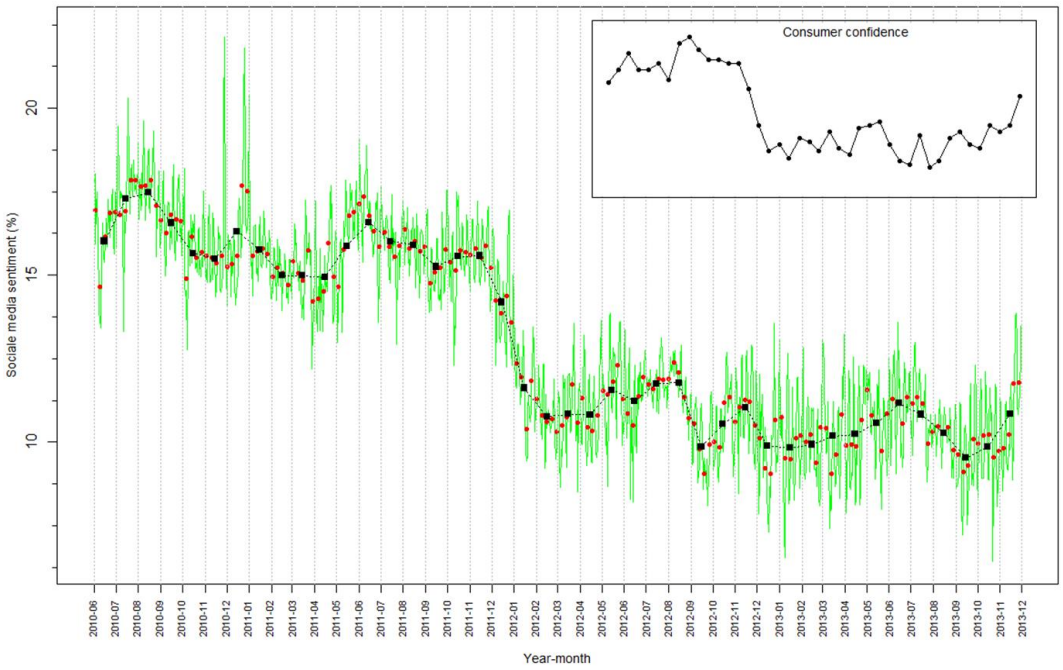


Figure 3.4-1
Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. In the insert the development of consumer confidence is shown for the same period.



Thus it might become possible to produce useful weekly sentiment indicators, even on the first working day after the week studied. Granger causality studies revealed that it is more likely that changes in consumer confidence precede those in social media sentiment than vice-versa. A comparison of the development of various seven-day sentiment aggregates with the monthly consumer confidence series confirmed this finding and revealed that the social media sentiment lag is most likely in the order of seven days. These and other research findings are consistent with the notion that changes in consumer confidence and social media sentiment are affected by an identical underlying phenomenon. An explanation for this phenomenon can be found in the Appraisal-Tendency Framework (Han et al., 2007), which is concerned with consumer decision-making. In this framework, it is claimed that a consumer decision is influenced by two kinds of emotions, namely the incidental and the integral. In this framework, the

integral emotion is relevant for the decision at stake, whereas the incidental emotion is not. Based on this theory, consumer confidence is likely to be influenced mainly by the incidental emotion, as consumer confidence is also not measured in relation to an actual decision to buy something. This suggests that the sentiment in social media messages might reflect the incidental emotion in that part of the population that is active on social media. Because of the general nature of the latter, one could denote this the "mood" of the nation in the context of consumer decision-making. More detailed results of this study are described in Daas and Puts (2014b).

4. Concluding remarks

The arrival of Big Data presents new opportunities for official statistics, but also poses new challenges. Important challenges for official statistics are: i) dealing with the selectivity of Big Data, ii) editing data at large scale and iii) reducing the volume of the data without losing (too much) information. Since there are different types of Big Data sources, e.g. human-generated, sensor-, and transaction-based, each source should be studied and judged on its own merits. Most important is to work data-driven while letting go of the sample-oriented perspective so familiar to statisticians (Daas and Puts, 2014a); a data-driven approach may prevent that any interesting findings are judged against the wrong framework. At this point in time, the area of Big Data research is just emerging and 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 (Bethlehem, 2009). 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 challenges identified above will need to be addressed. In particular there is a need for new legislation, statisticians with a new skill and mind set ('data scientists'), new methods and appropriate computational facilities (London Workshop, 2014; ASA-working group, 2014). The work would also benefit from an intensified international cooperation between data providers, scientists and official statisticians.

Acknowledgements

The authors gratefully acknowledge their Statistical Netherlands colleagues Edwin de Jonge, Joep Burger, Bart Buelens, Jan van den Brakel, May Offermans, Barteld Braaksma and Peter Struijs for stimulating discussions and constructive remarks. This work would not have been possible without the support of the innovation programme at Statistics Netherlands.

References

ASA-working group (2014), "*Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society*", report of a Working Group of the American Statistical Association, Located at: http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf

Bethlehem, J. G. (2009), "*The rise of survey sampling*", Statistics Netherlands Discussion Paper 09015, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Buelens, B., Daas, P., Burger, J., Puts, M., and Van den Brakel, J. (2014), "*Selectivity of Big Data*", Discussion Paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P. J. H., and Burger, J. (2014), "*Profiling Big Data sources to assess their selectivity*", abstract for the New Techniques and Technologies for Statistics conference 2015, Brussels, Belgium.

Daas, P. J. H., and Puts, M. J. H. (2014a), "Big Data as a source of statistical information", *The Survey Statistician*, 69, pp. 22–31.

Daas, P. J. H., and Puts, M. J. H. (2014b), "*Social Media Sentiment and Consumer Confidence*", European Central Bank Statistics Paper Series, 5, Frankfurt, Germany.

Daas, P., Puts, M., Ossen, S., and Tennekes, M. (2014), "*Processing and methods for Big Data: a traffic index based on huge amounts of road sensor data*", paper presented at the Conference of European Statistics Stakeholders, Rome, Italy.

Daas, P. J. H., Roos, M., Van de Ven, M., and Neroni, J. (2012), "*Twitter as a potential data source for statistics*", Discussion Paper 201221, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P. J. H., and Van der Loo, M. P. J. (2013), "*Big Data (and official statistics)*", paper presented at the 2013 Meeting on the Management of Statistical Information Systems, Paris–Bangkok, France-Thailand, Located at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

Fan, J., Han, F., and Liu, H. (2014), "Challenges of Big data analysis", *National Science Review*, 1, pp. 293-314.

Fyhrlund, A., Fridlund, B., and Sundgren, B. (2005), "Using Text Mining in Official Statistics", *Knowledge Mining, Proceedings of the NEMIS 2004 Final Conference, Studies in Fuzziness and Soft Computing, 185*, pp. 201-211.

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., and Khan, A. (2013), "*What does "Big Data" mean for Official Statistics?*", paper for the High-Level Group for the Modernization of Statistical Production and Services, Located at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614.

Hager, G., and Wellein, G. (2010), *Introduction to High Performance Computing for Scientists and Engineers*, Boca Raton: Chapman & Hall/CRC Computational Science.

Han, S., Lerner, J. S., and Keltner, D. (2007), "Feelings and consumer decision making: the appraisal-tendency framework", *Journal of Consumer Psychology*, 17, pp. 158–168.

Heerschap, N. M., Ortega Azurduy, S. A., Priem, A. H., and Offermans, M. P. W. (2014), "*Innovation of tourism statistics through the use of new Big Data sources*", paper presented at the Global Forum on Tourism Statistics, Prague, Located at: http://www.tsf2014prague.cz/assets/downloads/Paper%201.2_Nicolaes%20Heerschap_NL.pdf

London Workshop (2014), "*Statistics and Science*", report on the London Workshop on the Future of the Statistical Sciences, Located at: http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

Nguyen, D-P., Gravel, R., Trieschnigg, R.B., and Meder, T. (2013), TweetGenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter*, 4, pp. 4-9.

Puts, M., Tennekes, M., and Daas, P. (2014), "*Using Road Sensor Data for Official Statistics: Towards a Big Data Methodology*", paper presented at the Strata 2014 Conference, Barcelona, Spain.

Saporta, G. (2000), "*Data Mining and Official Statistics*", paper presented at Quinta Conferenza Nationale di Statistica, Rome, Italy.

Silver, N. (2012), "*The Signal and the Noise: Why So Many Predictions Fail —but Some Don't*", New York: Penguin Group.

Struijs, P., Braaksma, B., and Daas, P. (2014), "Official Statistics and Big Data", *Big Data & Society, April–June*, pp. 1–6.

Tennekes, M., and Offermans, M. P. W. (2014), "*Daytime population estimations based on mobile phone metadata*", paper presented at the Joint Statistical Meetings 2014, Boston, USA.