DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 249–262, http://dx.doi.org/10.1515/JOS-2015-0016

Big Data as a Source for Official Statistics

Piet J.H. Daas[1], *Marco J. Puts*[1], *Bart Buelens*[1], *and Paul A.M. van den Hurk*[1]

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 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. This article discusses the exploration of both opportunities and challenges for official statistics associated with the application of Big Data. Experiences gained with analyses of large amounts of Dutch traffic loop detection records and Dutch social media messages are described to illustrate the topics characteristic of the statistical analysis and use of Big Data.

Key words: Large data sets; traffic data; social media.

1. Introduction

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 introduction of the term 'Big Data' (Lynch 2008). Big Data sources can generally be described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making". This definition is a variant of the definition proposed by Gartner (Beyer and Douglas 2012). For more general information on Big Data and their innovative potential, the reader is referred to Manyika et al. (2011).

In addition to generating new commercial opportunities in the private sector, Big Data are potentially a very interesting data source for official statistics, either for use on their own, or in combination with more traditional data sources such as sample surveys and administrative registers (Cheung 2012). However, extracting relevant and reliable information from Big Data sources and incorporating it into the statistical production process is not an easy task (Daas et al. 2012a). Importantly, the statistical point of view has been underexposed in the work that has been "published" on Big Data so far; this work has been published mainly on weblogs and in conference and white papers. The majority of these publications have an IT perspective as they predominantly focus on soft- and hardware issues, and largely fail to address important statistical issues such as coverage, representativity, quality, accuracy and precision. If Big Data are to be used for official

[1] Statistics Netherlands, Division of Process development, IT and methodology P.O. Box 4481, 6401 CZ, Heerlen, The Netherlands. Emails: pjh.daas@cbs.nl (corresponding author), m.puts@cbs.nl, b.buelens@cbs.nl, and pamvandenhurk@gmail.com

statistics, it is essential that these issues are considered and adequately dealt with (Cheung 2012; Daas et al. 2012a; Glasson et al. 2013; Groves 2011).

In this article we provide an overview of the current state of the research on the usage of Big Data for official statistics at Statistics Netherlands and the lessons learned so far. In the next section a description of two Big Data case studies is given, followed by a more general methodological discussion in Section 3. Finally, conclusions are drawn in Section 4.

2. Big Data Case Studies

In this section we report on two Big Data case studies conducted at Statistics Netherlands. These studies serve as examples and allow for a more general formulation of the statistical issues and challenges involved with the application of Big Data in official statistics. All analyses were performed with the open-source software environment R (R Development Core Team 2012) on a Fujitsu Celsius M470-2 workstation with a 64-bit Windows 7 operating system, 32GB of RAM, 512 GB solid state drive and a 1 TB hard disk. Data were imported into R from CSV files which usually each contained one million rows of data. Each file was subsequently processed and analysed. Results were stored as CSV files. This approach was fast and flexible and sufficed for the studies described in this article.

2.1. Analysis of Traffic Loop Detection Data

Traffic loop detection data consist of measurements of traffic intensity. Each loop counts the number of vehicles per minute that pass at that location, and measures speed and vehicle length one. Such data are interesting for traffic and transport statistics and potentially also for statistics on other economic phenomena related to transport. On the particular day studied, data were collected at 12,622 measurement locations on Dutch roads. The data are stored centrally in the National Data Warehouse for Traffic Information (NDW) and managed by a collaboration of participating government organizations (NDW 2012). The National Data Warehouse contains historic traffic data collected from 2010 onwards. To determine the usability of the NDW data for statistics and to get an idea of its peculiar features, we started by studying minute-level data for all locations in the Netherlands for a single day: December 1st, 2011. The data set extracted from the NDW contained 76 million records, one million per CSV file, which were imported into R via the LaF package (Van der Laan 2013). This package supports loading the data in blocks, enabling the processing of enormous amounts of data without fitting all the data into memory.

Data were first aggregated over all loops, resulting in a series of total counts of all vehicles in the Netherlands at minute intervals. The change of this total count through the day is shown in Figure 1A. The overall profile displays clear morning and evening rush hour peaks around 8 am and 5 pm respectively. Importantly, however, there is a huge variation in the numbers of vehicles detected in subsequent minutes. This phenomenon is caused by the fact that – for a substantive number of minutes – data were only available for a subset of all detection loops in the country. This appeared to be caused by some computers failing to submit data to the warehouse at certain time points.

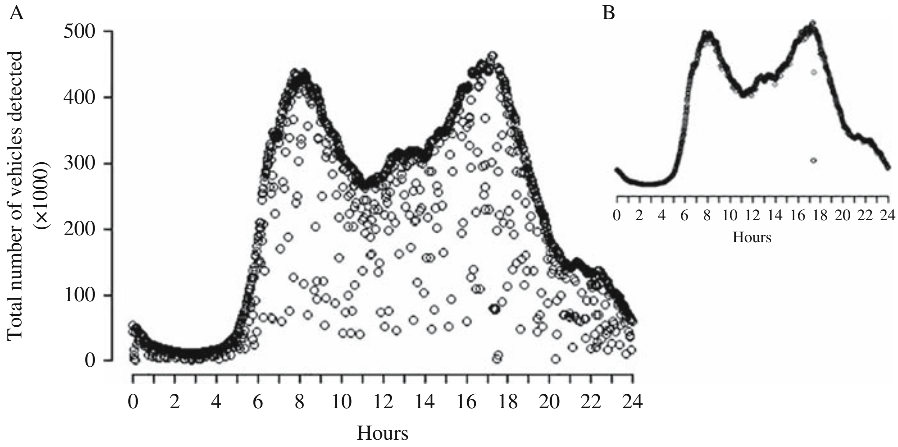

Fig. 1. (A) Total number of vehicles detected per minute in the Netherlands on December 1st, 2011. (B) Results after correcting for missing data.

From a statistical point of view there are various ways to solve such a missing data problem (De Waal et al. 2011).

Because aggregated data were used and this was our first experience with huge amounts of data, we opted for the simplest solution: add (impute) data reported by the same location during a short interval before or after the time point of missing data (if available). More specifically, a sliding, symmetrical five-minute time window to impute data at missing time points for the entire data set was applied. The resulting data pattern is shown in Figure 1B. Except for a period shortly after 5 pm, the majority of the missing data points were adequately replaced with timely data of the same measurement location. As a result of this data-editing procedure a total of nearly 35.8 million vehicle counts were added, which is slightly more than twelve percent of the number of vehicles originally counted, 294.7 million. Alternative model-based approaches can be applied and are preferred when traffic loop data are studied for smaller areas (more on this below).

The edited data set was used to create maps that indicate the number of vehicles for each measurement location for each time point by means of colour coding. Next, by sequencing these maps, a movie was created that displays the changes in vehicle counts for all locations during the day. Thus, this movie (not shown here) illustrates the increases and decreases in traffic intensity in the Netherlands throughout the day studied (Daas et al. 2012b). Unsurprisingly, the traffic intensity between the four major cities in the Netherlands (Amsterdam, Rotterdam, Utrecht and The Hague) was especially high, during all working hours and in the early evening.

Besides the total number of vehicles, the number of vehicles in various length categories was also studied. Because not all detection locations are able to differentiate between different vehicle lengths, only those that are able to do so were used. This subset consisted of 6,002 detection locations, which represented 48 percent of the total number of locations. Vehicles were sorted into three length categories: small (<= 5.6 metre), medium-sized (> 5.6 and <= 12.2 metre), and large (> 12.2 metre). Again, the imputed data set was used. Because the small vehicle category comprised around 75 percent of all vehicles detected, as compared to twelve percent for the medium-sized and 13 percent for the large

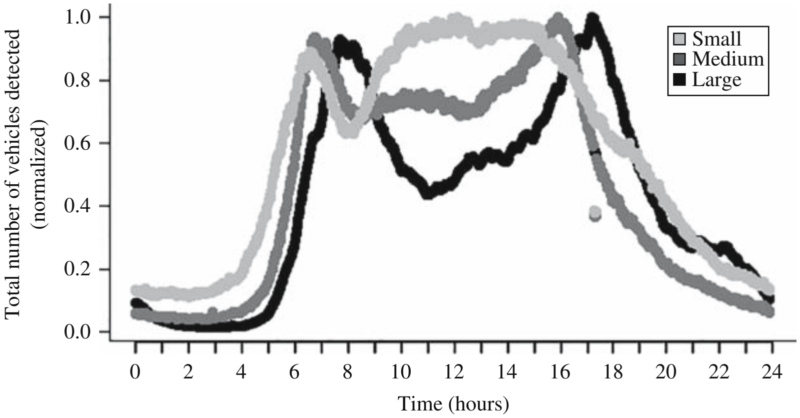

Fig. 2. Normalized numbers of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Numbers of small (<= 5.6 meter), medium-sized (> 5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown. Profiles are normalized by dividing by the maximum value of each series to more clearly reveal the differences. Maximum values are 119,523, 8,673 and 8,599 for small, medium and large vehicles, respectively.

vehicles categories, the normalized results for each category are shown in Figure 2. This figure illustrates the difference in driving behaviour between the three vehicle length categories. The small vehicle category displays clear morning and evening rush-hour peaks at 8 am and 5 pm respectively, in line with the overall profile described above (Figure 1). This finding is not unexpected, as this category of vehicles constitutes the vast majority of all vehicles. The medium-sized vehicles in turn have both an earlier morning and evening rush-hour peak, at 7 am and 4 pm respectively. Finally, the large vehicle category shows a clear morning rush-hour peak around 7 am and more dispersed driving behaviour during the remainder of the day; after 3 pm the number of large vehicles gradually declines without any apparent evening rush-hour peak. Most remarkable is the decrease in the relative number of medium-sized and large vehicles detected at 8 am, that is, during the morning rush-hour peak of the small vehicles. This may be caused by a deliberate attempt of the drivers of the medium-sized and large vehicles to avoid the morning rush-hour peak of the small vehicles or an effect of the more intense traffic (of small vehicles) around that time. Considering these differences, differentiation between vehicles of various lengths when creating a traffic index would not only enable more granular traffic statistics but can also provide more detailed information on transport and phenomena related to economic growth.

In addition to the analysis of traffic intensity at an aggregated level across all detection loops, the traffic intensity profile of a number of individual measurement locations was also studied, for example on highway A4 near Bergen op Zoom. The total number of vehicles detected at this location is shown in Figure 3. Detection at this location displays the same rush-hour peaks as in Figure 1. In addition, the characteristic volatile behaviour of traffic intensity data at the microlevel is shown. Given that this detection location does not suffer from missing data, the changes in the number of vehicles counted each minute are the result of real changes in the number of vehicles passing at this location. However, these rapid fluctuations are not very informative for the production of a traffic index,

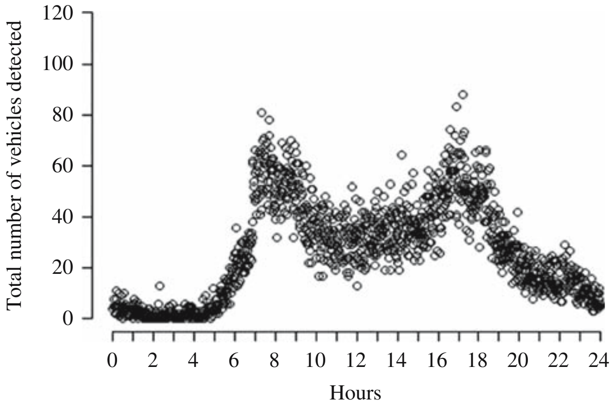

Fig. 3. Total number of vehicles counted by a detection location on highway A4 near Bergen op Zoom.

as interest is focused more upon gradual, long-period, changes, for example, weekly or monthly changes in the number of vehicles (of a certain length class) in a specific region. We are currently studying statistical modelling methods that can deal adequately with these kinds of Poisson-distributed data, such as Bayesian-based signal filters (Manton et al. 1999). These methods need to be applied in a reasonable time to the large amounts of loop detection data. The latter requires high-performance computing techniques (NAS 2013) when applied to the data of all loops in the whole country.

In the analyses in this section, we have assumed that all measurements are without error, except when entire records are missing. The missing records have been imputed using fixed values, not taking into account uncertainty associated with the imputation procedure. Alternatively, a multiple imputation approach could be implemented to account for such uncertainty, which would result in variances and confidence intervals for the aggregates shown above. The aggregates are obtained simply by summing individual loop counts. We have not conducted any form of inference or estimation (except for the imputations). In the future we may do so in order to obtain estimates that are representative of all Dutch highways, including those without traffic loops. A predictive modelling approach would need to be developed, resulting in estimated counts at locations without loops. This would lead to estimated aggregates and variance estimates reflecting the uncertainty of the estimation procedure.

2.2. *Analysis of Social Media Messages*

It is estimated that around 70 percent of the Dutch population actively posts messages on social media (Eurostat 2012). The three million or so Dutch messages generated each day (Daas and Puts 2014) may be an interesting data source for official statistics because they reflect many different aspects of our daily lives. We have studied two aspects of social media messages: content and sentiment. Studies of the content of Dutch Twitter messages – the dominant publicly available social medium in the Netherlands (see below) – revealed that nearly 50 percent of the messages are “pointless babble” (Daas et al. 2012a). In the remainder of the messages, spare-time activities, work, media (TV & radio) and politics were predominantly discussed. This finding suggests that these messages could be used to extract opinions, attitudes, and sentiments towards these topics, opening up possibilities to collect a considerable amount of interesting information quickly without any response burden. The major problem in analysing social media messages is discriminating the informative from the noninformative ones. Because of the large share of the noninformative "babble" messages, usage of the more serious (informative) messages is negatively affected as many words of interest occur in both types of messages. Text mining approaches to automatically differentiate between both groups of messages have not been very successful so far (Daas et al. 2012a).

Another potential source of information in social media messages is their sentiment. Access to over 1.6 billion public messages written in Dutch from a large number of social media sites was obtained using an infrastructure provided by Coosto (2013). Public messages were sourced from the largest social media sites used by Dutch individuals, such as Twitter, Facebook, Hyves, Google+, and LinkedIn, as well as from numerous public Dutch weblogs and forums. The overall profile of the number of messages created per day revealed that from June 2010 onwards, increasing numbers of messages were generated in the Netherlands on a daily basis. The latter date corresponds to the period during which Coosto started to include huge numbers of Twitter messages in their Hadoop-based distributed database. We therefore used June 2010 as the starting date for our studies, with August 2012 as the end date. Messages could be selected from the database with a query language and a secure web interface. Coosto also determined the sentiment of each message by counting the number of positive and negative words following the general approach described in Golder and Macy (2011). Messages were classified as positive, negative or neutral depending on their overall score. A more detailed description of this part of the work can be found in Daas and Puts (2014).

Since several studies have been performed in English-speaking countries attempting to link the sentiment in social media to consumer confidence (O'Connor et al. 2010; Lansdall-Welfare et al. 2012) we were interested in studying this "relation" for the Netherlands. We looked at the sentiment in messages produced on the various platforms covered by the Coosto data set. The results were intriguing. The development of the sentiment in all Facebook messages produced during the period studied, nearly 170 million (almost ten percent of all messages produced), was found to correlate highly with consumer confidence; $r = 0.84$. Combing the sentiment of all Facebook and Twitter messages, slightly over 1.4 billion (close to 90% of all messages), with a linear model increased the correlation to $r = 0.88$. To reduce the risk of discovering spurious or false correlations, the series were additionally checked for cointegration. Cointegration provides a stronger argument as it checks for a common stochastic drift, indicating that series exhibit fluctuations around a common trend (Engel and Granger 1987). Here, it was found that the sentiment in Facebook and the combination of Facebook and Twitter both cointegrated with consumer confidence, suggesting a strong association between the developments in both series. Remarkably, the sentiment in Twitter messages only correlated less, $r = 0.61$, and did not cointegrate.

Figure 4 displays the survey-based Consumer Confidence series (Statistics Netherlands 2013) and the corresponding Dutch social media sentiment findings for the period studied. Both series relate quite well. This association is remarkable, as the

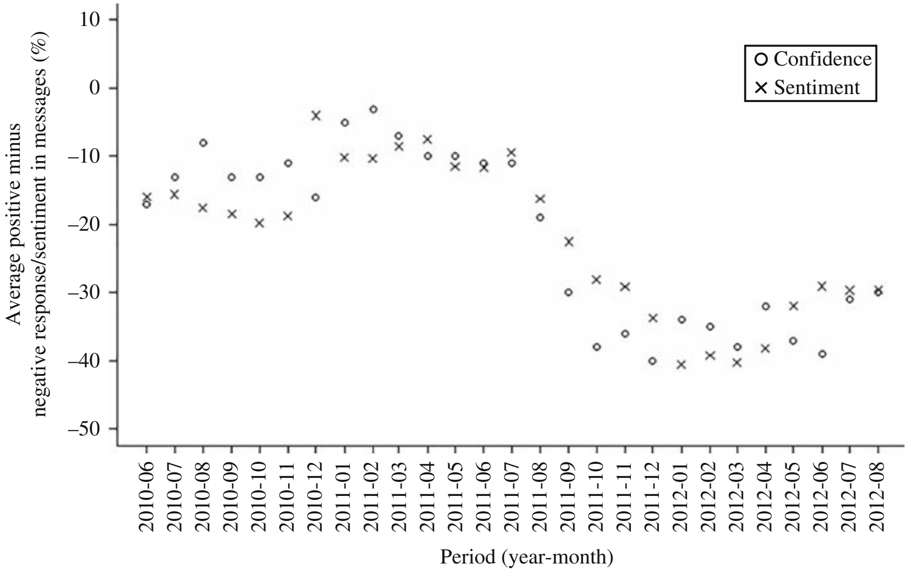

Fig. 4. Comparison of Dutch consumer confidence (○) and the sentiment in Dutch Facebook and Twitter messages on a monthly basis (×). A correlation coefficient of r = 0.88 is found for both series.

populations from which the data are obtained are very different. Dutch consumer confidence is obtained from a random sample from the population register, with around 1,000 persons responding each month. Due to sampling variance, the standard errors of the consumer confidence series shown in Figure 4 are on average approximately 2.0. The sentiment in Dutch Facebook and Twitter messages is derived from around 52 million *messages* generated each month. These messages are created by a considerable part of the population, 70 percent according to Eurostat (2012), but i) not all social media messages created in the Netherlands are written in Dutch and ii) different users post varying numbers of messages on various platforms. We have not attempted to estimate the sentiment of the (unknown) subpopulation who does not contribute to social media platforms. Consequently, our social media sentiment series is not subject to sampling, modelling or prediction uncertainty, but may be biased because of differences in the composition of the Dutch population and those active on social media. Previous work by Daas et al. (2012a) also revealed that the number of Twitter messages can vary from 200 per day to not even one message a month for a single person. More recent work has confirmed that the association between both series remains stable over time and that consumer confidence and social media sentiment are related from a Granger-causality perspective (more in Daas and Puts 2014).

3. Discussion

The two case studies described in this article reveal several issues that need to be addressed before Big Data can become a useful and reliable data source for the field of official statistics. These issues, the most important considerations, the way we have dealt with them and the lessons learned are discussed below.

3.1. Data Exploration

Typically, Big Data sets are made available to us, rather than designed by us. As a consequence, their contents and structure need to be understood prior to using the data for statistical analysis (Hassani et al. 2014). This first step is called data exploration, which is aimed at revealing data structure and patterns and, no less important, at assessing the quality of the data as revealed by the presence of errors, anomalies and missing data. Visualisation methods have been proven to be very insightful for such tasks (Fry 2008; Zikopoulos et al. 2012, ch. 7). Recently, certain visualisation methods have been developed that are particularly suited to the exploration of Big Data. Examples are tableplots (Tennekes et al. 2013) to display Big Data with many variables and 3D heatmaps to study variability in multivariate continuous data (Daas et al. 2012b). Sequencing 2D plots into animations is useful to visualise temporal and other aspects of Big Data (Daas et al. 2012b).

3.2. Missing Data

By studying the traffic intensity data on a minute-by-minute level, we discovered that part of the data were missing. If we had analysed the data aggregated at hourly or daily levels, we would have reduced the amount of data studied but would not have noticed that missing data is such a big problem. Since Statistics Netherlands plans to use NDW data to produce reliable traffic and transport intensity statistics at a detailed level, the missing data problem needs to be solved. Missing data is not a problem unique to the traffic loop data set, as other data sources are susceptible to missing data too. For instance, server downtime and network outages can lead to missing social media messages or mobile phone data. However, in the end, the time spent on processing also needs to be reduced to a manageable level to enable the production of frequent statistics. Currently statistical models are being explored that are able to cope adequately with missing data and can be applied to enormous amounts of data in a reasonable amount of time. For such an approach to be successful, the combination of the IT infrastructure available and the ease with which a modelling method can be upscaled needs to be assessed (NAS 2013). We are currently focusing on Bayesian approaches as these are applied to enormous amounts of data in other areas of science and are well suited to capturing various forms of uncertainty. The high-performance computing needs can be met at relatively low cost by using the large amounts of computing power provided by the graphics processing units available on many modern graphics cards (Scott et al. 2013).

3.3. Volatility

The number of vehicles detected by individual loops fluctuates considerably from minute to minute. These fluctuations are caused by real changes in the number of vehicles detected but are not very informative from a statistical point of view as they occur at too high a time resolution. Similarly, sentiment analyses on a daily and weekly basis suffer from a volatility that is not seen at monthly intervals (Daas and Puts 2014; O'Connor et al. 2010). It is therefore recommended to develop statistical methods able to cope with volatile behaviour. Possible methods under consideration are the application of moving averages and advanced filtering techniques (e.g., a Kalman filter or time-series modelling).

3.4. Selectivity

The analyses described in Section 2 apply to traffic intensity on roads equipped with traffic loop sensors, and to the sentiment analysis of people who post Dutch Facebook or Twitter messages on social media websites. It is important to realize that both data sets are created by only a subset of the total population in the Netherlands: only vehicles driving on the major Dutch roads were counted and only the sentiment of a subset of all people in the Netherlands was probed, respectively. The subpopulations from which these Big Data sources were derived are not typical target populations for official statistics. Therefore the data are likely to be selective and not representative of a target population of interest. In addition, both sources contain data resulting from the registration of events. These are vehicles passing and messages sent respectively. Both lack directly available data on the units of interest. Usually, the representativity of Big Data can be assessed through the careful comparison of characteristics of the covered population and the target population. Unfortunately, this may prove problematic for these sources, as hardly any such characteristics are available to conduct such a comparison (Buelens et al. 2014). For instance, vehicles can not be uniquely identified in the traffic loop data as licence plate data are absent. Little is known about the people posting on social media; often only their username is known but not their age or gender. In situations where at least some background information is available, the selectivity issue can be assessed and probably resolved. Alternatively, profiling approaches could be used to extract features to estimate, for instance, the chance that a user is male or female (Flekova and Gurevych 2013). Perhaps this could be achieved through predictive modelling, using a wide variety of algorithms known from statistical learning and data mining techniques (Hastie et al. 2009). These are modelling methods not traditionally used in official statistics. Buelens et al. (2012) explore some possibilities for applications of data mining methods in official statistics. More on this topic can be found in ASA (2014).

3.5. Legal Considerations

Privacy and security are issues that may impede NSIs' use of Big Data. In contrast to the legal basis that permits the use of administrative data sources by a lot of NSIs, the use of privately owned Big Data, such as mobile phone data, needs to be specifically arranged (De Jonge et al. 2012). But even for publicly accessible data, such as price and product information on websites, questions of ownership and purpose of publication can be raised. And even if there are no legal impediments, public perception is a factor that must be taken into account. These concerns have to be taken seriously and tackled one at a time. Fortunately, there are measures that can be taken to overcome at least some of the obstacles, for example, by anonymizing unique identifiers, removing the privacy-sensitive part of a Global Positioning System track (e.g., the first and last 100 metres) or by using informed consent. If a reduction of response burden can be offered, this can be very helpful, also in getting the support of the general public. In the long run, changes in legislation may be considered, to ensure continuous data access for official statistics. But it remains important to stay in line with public opinion, because credibility and public trust are important assets. Within the European Union, changes in European legislation must also be considered. In addition to national laws, European laws or regulations can impede the collection of data, even if the current Dutch legislation does not present any problem (more in Struijs and Daas 2013).

3.6. Data Management

Long-term stability may also 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. The Big Data sources encountered so far seem subject to frequent modifications, possibly limiting their long-term use. This suggests a need for more flexible data processing and evaluation strategies, which will have to put more emphasis on ongoing data and metadata management to identify, describe and re-evaluate new sources. Data management is also affected by data ownership, copyright and the purpose for which data are registered. Privately owned Big Data in particular may need special arrangements and will probably also incur costs. The two data sources discussed in this article are examples of each. The organisation that maintains traffic loop data is funded by the Dutch government which means that, as laid down in the Statistics Netherlands Act, they provide us with the data free of charge. For the study of social media data, however, costs were incurred, as the data are collected by a privately owned company.

3.7. High-Performance Computing

Processing enormous amounts of data within a reasonable amount of time requires dedicated and specialized computing infrastructures. Hence it can be expected that the inclusion of Big Data as a source for official statistics will certainly affect the IT environment of NSIs. Our experiences so far, however, reveal that considerable progress can be made even with a limited budget. Having a secure computer environment with many fast processors, large amounts of RAM and fast disk access certainly helps. Several important considerations are described in NAS (2013) and Schutt and O'Neil (2013). Parallel processing is the way to speed things up. For instance, we have found that processing traffic loop data in parallel in R results in a 17x speedup over the original (serial) processing time. We are currently using (multicore) general-purpose computing on graphics processing units and are looking at distributed computing, such as our own secure local cluster.

3.8. New Skills Needed

In order to work with Big Data specific technical expertise is needed, such as knowledge of advanced (high-performance) computing and data engineering. These skills speed up the ease with which Big Data can be incorporated into the statistical process and the way it is analysed. In our office, Big Data is usually processed with R or Python. Besides knowing the language, the most important skill here is knowing how to write a program that is able to access and analyse all the data within a reasonable amount of time. Several of our colleagues have written R packages specifically devoted to these tasks, such as LaF (Van der Laan 2013) and ffbase (De Jonge et al. 2014). Moreover, the models used for Big Data must be able to address the levels of complexity that huge data sets can reveal. This makes many of the standard approaches used in official statistics limited in utility and performance (NAS 2013). The algorithmic-oriented models developed in fields outside statistics might be more applicable here (Breiman 2001; Hastie et al. 2009).

Perhaps just as important is the attitude of the people involved. Working with Big Data requires an open mindset and the ability not to see all problems *a priori* in terms of sampling theory, as Big Data are more similar to large sets of observational data (Daas and Puts 2014). The term "data scientist" has been coined for researchers with the skills identified above (Schutt and O'Neill 2013). We have solved the initial need by including experimentally trained researchers in our Big Data efforts, as they are more practically oriented and are more accustomed to deriving theory from data. The strict difference between methodology, software engineering and IT hardware expertise commonly used at NSIs is also becoming less well defined. At Statistics Netherlands, a group of data scientists is currently being formed. The work of this group is expected to be beneficial to many of the areas of official statistics, especially when large data sources and complex models are used.

4. Conclusions

The official statistics community can greatly benefit from the possibilities offered by Big Data. However, care is needed when trying to implement these sources in official statistics. The two Big Data case studies described show typical issues including missing data, volatility and selectivity, which all need to be adequately dealt with. For this reason, investment in specific research and skills development is needed. In addition, various new areas of expertise are considered necessary to fully exploit the information contained in Big Data. In particular, knowledge is required from the fields of mining and analysing massive data sets (Rajaraman and Ullman 2011; Hassani et al. 2014), high-performance computing (NAS 2013), and the new emerging discipline commonly referred to as "Data Science" (Schutt and O'Neill 2013). We expect to see some Dutch official statistics derived from Big Data in the coming years. When produced in a methodologically sound manner, official statistics based on Big Data can be cheaper, faster and more detailed than the official statistics known to date. For these endeavours to become successful, it is essential that they are supported by the general public and both Dutch and European legislation.

5. References

ASA. 2014. *Discovery With Data: Leveraging Statistics with Computer Science to Transform Science and Society*. July 2, 2014 version. Available at: http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf (accessed July 2014).

Beyer, M.A. and L. Douglas. 2012. *The Importance of 'Big Data': A Definition*. Gartner report, June version, ID Number: G00235055. Available at: http://www.gartner.com/it-glossary/big-data/ (accessed January 2013).

Breiman, L. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16: 99–231. Doi: http://dx.doi.org/10.1214/ss/1009213726.

Buelens, B., H.J. Boonstra, J. van den Brakel, and P. Daas. 2012. *Shifting Paradigms in Official Statistics: from Design-Based to Model-Based to Algorithmic Inference*. Discussion paper 201218, Statistics Netherlands, The Hague/Heerlen.

Buelens, B., P. Daas, J. Burger, M. Puts, and J. van den Brakel. 2014. *Selectivity of Big Data*. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Cheung, P. 2012. *Big Data, Official Statistics and Social Science Research: Emerging Data Challenges*. Presentation at the December 19th World Bank meeting, Washington. Available at: http://www.worldbank.org/wb/Big-data-pc-2012-12-12.pdf (accessed January 2013).

Coosto. 2013. Main page. Available at: http://www.coosto.com/uk/ (accessed August 2013).

Daas, P.J.H. and M.J.H. Puts. 2014. *Social Media Sentiment and Consumer Confidence*. Paper for the Workshop on using Big Data for Forecasting and Statistics, April 7–8, Frankfurt, Germany. Available at: https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf (accessed April 2015).

Daas, P.J.H., M. Roos, M. van de Ven, and J. Neroni. 2012a. *Twitter as a Potential Data Source for Statistics*. Discussion paper 201221, The Hague/Heerlen: Statistics Netherlands.

Daas, P., M. Tennekes, E. de Jonge, A. Priem, B. Buelens, M. van Pelt, and P. van den Hurk. 2012b. *Data Science and the Future of Statistics*. Presentation at the first Data Science NL meetup, Utrecht University, Utrecht. Available at: http://www.slideshare.net/pietdaas/data-science-and-the-future-of-statistics (accessed December 2012).

De Jonge, E., M. van Pelt, and M. Roos. 2012. *Time Patterns, Geospatial Clustering and Mobility Statistics Based on Mobile Phone Network Data*. Discussion paper 201214, The Hague/Heerlen: Statistics Netherlands.

De Jonge, E., J. Wijffels, and J. van der Laan. 2014. "ffbase: Basic Statistical Functions for Package ff. R package version 0.11.3." Available at: http://cran.r-project.org/web/packages/ffbase/index.html (accessed April 2015).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

Engle, R.F. and C.W.J. Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica* 55: 251–276.

Eurostat. 2012. *Internet Access and Use*. Eurostat newsrelease 185/2012, December 18, 2012. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF (accessed January 2013).

Flekova, L. and I. Gurevych. 2013. *Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media*. Paper for the evaluation lab on uncovering plagiarism, authorship, and social software misuse at Conference and Labs Evaluation Forum 2013, September 23–26, Valencia, Spain.

Fry, B. 2008. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. Sebastopol, CA: O'Reilly Media Inc.

Glasson, M., J. Trepanier, V. Patruno, P. Daas, M. Skaliotis, and A. Khan. 2013. *What does "Big Data" mean for Official Statistics?* Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.

Golder, S.A. and M.W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 30: 1878–1881. Doi: http://dx.doi.org/10.1126/science.1202775.

Groves, R.M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75: 861–871. Doi: http://dx.doi.org/10.1093/poq/nfr057.

Hassani, H., G. Saporta, and E. Sirimal Silvia. 2014. "Data Mining and Official Statistics: The Past, the Present and the Future." *Big Data* 2: 1–10. Doi: http://dx.doi.org/10.1089/big.2013.0038.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Science + Business Media, LLC.

Lansdall-Welfare, T., V. Lampos, and N. Cristianini. 2012. "Nowcasting the Mood of the Nation." *Significance* 9: 26–28. Available at: http://www.significancemagazine.org/details/magazine/2468761/Nowcasting-the-mood-of-the-nation.html (accessed January 2013).

Lynch, C. 2008. "Big Data: How Do Your Data Grow?" *Nature* 455: 28–29. Doi: http://dx.doi.org/10.1038/455028a.

Manton, J.H., V. Krishnamurthy, and R.J. Elliott. 1999. "Discrete Time Filters for Double Stochastic Poisson Processes and Other Exponential Noise Models." *International Journal of Adaptive Control and Signal Processing* 13: 393–416.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Report of the McKinsey Global Institute, McKinsey & Company.

NAS. 2013. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

NDW. 2012. *The Database Explained. Brochure of the National Data Warehouse for Traffic Information, March*. Available at: http://www.ndw.nu/download_files.php?action=download_file&file_hash=209140a807e959f06646b0311f79de26 (accessed December 2012).

O'Connor, B., R. Balasubramanyan, B.R. Routledge, and N.A. Smith. 2010. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*. Carnegie Mellon University, Research Showcase. Available at: www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf (accessed April 2015).

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rajaraman, A. and J.D. Ullman. 2011. *Mining of Massive Datasets*. Cambridge: Cambridge University Press.

Schutt, R. and C. O'Neil. 2013. *Doing Data Science: Straight Talk from the Frontline*. Sebastopol, CA: O'Reilly Media.

Scott, S.L., A.W. Blocker, F.V. Bonassi, H.A. Chipman, E.I. George, and R.E. McCulloch. 2013. *Bayes and Big Data: The Consensus Monte Carlo Algorithm*. Bayes 250. Available at: http://www.rob-mcculloch.org/some_papers_and_talks/papers/working/consensus-mc.pdf (accessed April 2015).

Statistics Netherlands. 2013. *Consumer Confidence Survey*. Available at: http://www.cbs.nl/en-GB/menu/methoden/dataverzameling/consumenten-conjunctuur-onderzoek-cco.htm (accessed April 2013).

Struijs, P. and P.J.H. Daas. 2013. *Big Data, Big Impact?* Paper for the Seminar on Statistical Data Collection, September 25–27, Geneva, Switzerland.

Tennekes, M., E. de Jonge, and P.J.H. Daas. 2013. "Visualizing and Inspecting Large Datasets with Tableplots." *Journal of Data Science* 11: 43–58.

Van der Laan, J. 2013. *LaF: Fast Access to Large ASCII files*. R package version 0.5.

Zikopoulos, P., D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. 2012. *Harness the Power of Big Data*. New York: McGraw-Hill.

Received August 2013
Revised August 2014
Accepted September 2014