Big Data and Official Statistics

Daas P.J.H.[1], Puts M.J.[2], Buelens B.[3] and van den Hurk P.A.M.[4]
[1] Statistics Netherlands, Methodology sector, e-mail: pjh.daas@cbs.nl
[2] Statistics Netherlands, Traffic and transport sector, e-mail: m.puts@cbs.nl
[3] Statistics Netherlands, Methodology sector, e-mail: b.buelens@cbs.nl
[4] Statistics Netherlands, Methodology sector, e-mail: pam.vandenhurk@cbs.nl

Abstract

More and more data are being produced by an increasing number of electronic devices surrounding us and on the internet. The amount of data and the frequency at which they are produced are so vast that they are usually referred to as 'Big Data'. Because of their abundance and the fact that they reflect part of our daily lives, Big Data sources are very interesting from an official statistics point of view. This paper discusses opportunities and challenges associated with using Big Data for official statistics. Experiences obtained with analyses of large amounts of Dutch traffic loop detection records and Dutch social media messages are used to illustrate the topics specific to the statistical analysis of Big Data.

Keywords: Large data sets, Traffic data, Social media

Acknowledgements: The research described in this paper would not have been possible without the support of many people. The authors particularly wish to express their gratitude to Martijn Tennekes, Edwin de Jonge, Barteld Braaksma, Floris van Ruth and Marko Roos for their stimulating discussions on Big Data and official statistics.

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). These 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". This is a variant of the definition proposed by Gartner (2012). More on Big Data and their innovative potential can be found in the McKinsey report (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, harvesting the information from Big Data sources and incorporating it into the statistical production process is not an easy task (Daas 2012). The statistical point of view is remarkably underexposed in the work 'published' –mainly in weblogs, conference and white papers– on Big Data. The majority of these papers 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. When Big Data are being used for official statistics it is essential that these issues are considered and dealt with (Cheung 2012, Daas et al. 2012a, Groves 2011).

In this paper we provide an overview of the current state of the research on the use of Big Data for official statistics at our office and the lessons learned so far. Section 2 presents two Big Data case studies, followed by a methodological discussion in section 3. 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, see section 3.

2.1. Analysis of Traffic loop detection data

Traffic loop detection data consists of measurements of traffic intensity. Each loop counts the number of vehicles per minute that pass at that location, and measures speed and length. Such data are interesting for traffic and transport statistics and potentially also for statistics on other economic phenomena related to transport. To this date, data are collected at 12,622 measurement locations on Dutch roads and this number is still growing vastly. The data are centrally stored 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 peculiar features, we started by studying minute level data for all locations in the Netherlands for a single day: December 1st, 2011. The dataset extracted from the NDW contained 76 million records which were analysed in the open source software R environment (R Development Core Team 2012).

The data are aggregated over the loops, resulting in a series of counts at minute intervals. This series is shown in Figure 1A. The overall profile displays clear morning and evening rush hour peaks around 8 am and 5 pm respectively. Remarkably, there is huge variation in the numbers of vehicles detected in subsequent minutes: often a high count one minute decreases dramatically the next. This phenomenon is caused by the fact that –for quite some minutes– not all data are available for all detection loops in the country. This is caused by some computers failing to submit data to the warehouse some of the time. From a statistical point of view there are various ways to solve this missing data problem.

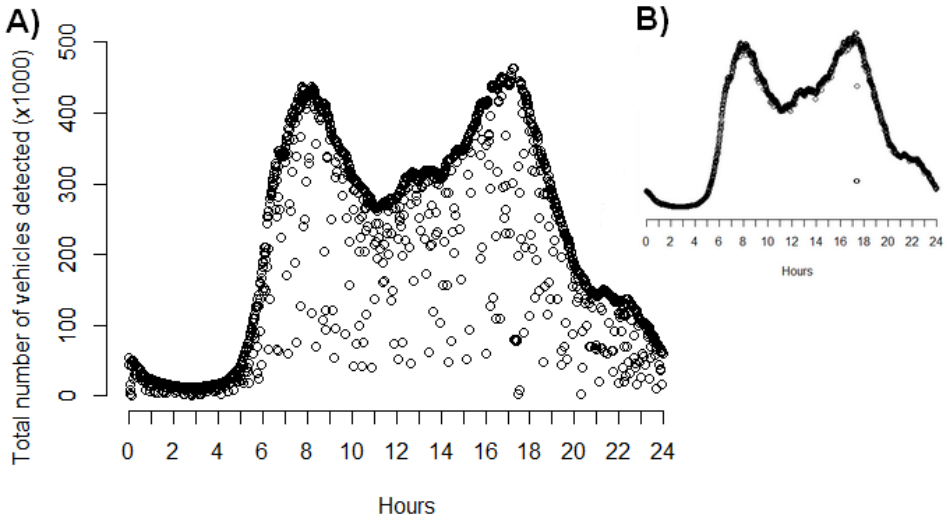

Figure 1. A) Total number of vehicles detected per minute in the Netherlands on December 1st, 2011. B) Results after correcting for missing data.

The simplest solution is to add (impute) the data for the absent detection sites in a particular minute, with data reported by the same locations during a short-time interval before or after that minute (if available). We did this for blocks of 5 minutes for the entire dataset. The result obtained is shown in Figure 1B. Except for a period shortly after 5 pm, the majority of the missing data issues were solved by this approach. As a result of imputation a total of 35,781,078 vehicle counts were added; this is a bit more than 12% of the number of vehicles originally counted, 294,702,822.

With the imputed dataset maps can be created indicating the number of vehicles for each measurement location (by using different colours) and, by combining these maps, a movie can be created that displays the changes in vehicle counts for all locations during the day. Such figures (not shown here) illustrate the increase and decrease of traffic in the Netherlands on the day studied (Daas et al. 2012b). Especially traffic between the four major cities in the Netherlands was and remained high during working hours and in the early evening.

Next the number of vehicles in various length categories was studied. Because not all detection sites are able to differentiate vehicle lengths, only those that are able to do so were used. This subset consisted of 6002 locations; 48% of the total number of unique locations. 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 in Figure 2. This illustrates the difference in driving behaviour. The small vehicles have clear morning and evening rush-hour peaks at 8 am and 5 pm respectively, in line with the overall profile (Figure 1). This is not unexpected as this category of vehicles contains the majority of all vehicles. 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

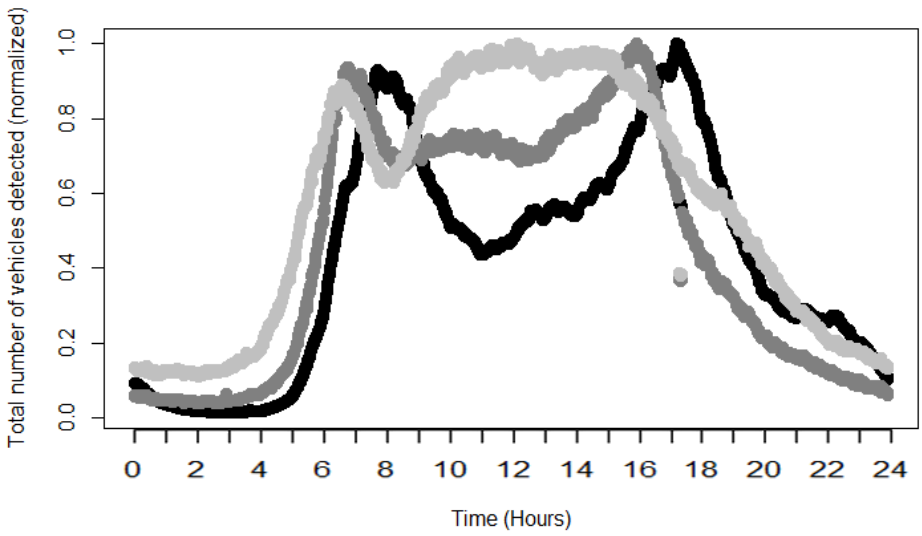

Figure 2. Normalized number of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Small (<= 5.6 meter), medium-sized (>5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown in black, dark grey and grey, respectively. Profiles are normalized to clearly reveal the differences in driving behaviour.

hour peak around 7 am and displays a more distributed driving behaviour during the remainder of the day. After 3 pm the number of large vehicles gradually declines. Most remarkable is the decrease in the relative number of medium-sized and large vehicles detected at 8 am, during the morning rush hour peak of the small vehicles. This may be caused by a deliberate action of the drivers of the medium-sized and large vehicles of wanting to avoid the morning rush hour peak of the small vehicles.

Apart from the macro-oriented approach, the 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 does not suffer from missing data and displays the same rush hour peaks as Figure 1. Figure 3 additionally illustrates the volatile behaviour of vehicle detection data at the micro level. The changes in the number of vehicles counted each minute are the result of real changes in the number of vehicles passing at the location. These changes are however, from a statistical point of view, not very informative and can even negatively affect analysis. For instance when the number of vehicles (of a certain length class) are studied at a detailed (regional) level. It is therefore recommended to develop statistical methods that cope with such volatility in measured data.

2.2. Analysis of Social media messages

It is estimated that around 70% of the Dutch population actively posts messages on Social media (Eurostat 2012). The millions of Dutch messages generated each day (Coosto 2012) may be an interesting data source for statistics. We studied social media messages

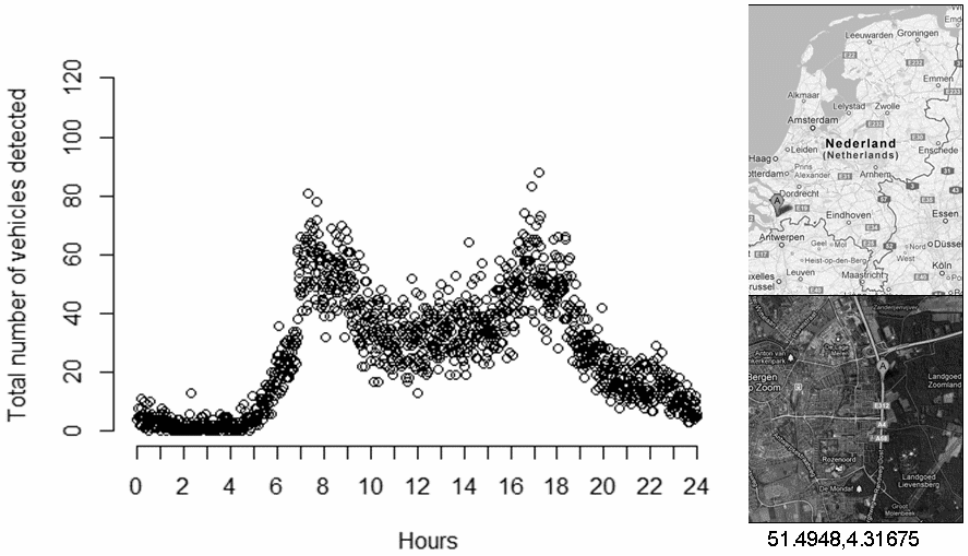

Figure 3. Total number of vehicles detected by a measurement location on highway A4 near Bergen op Zoom. The exact location is shown in the inserts.

from two points of view: content and sentiment. Studies of the content of Dutch Twitter messages –the dominant social medium in the Netherlands (see below)– revealed that nearly 50% of the messages were composed of 'pointless babble' (Daas et al. 2012a). In the remainder spare time activities, work, media (TV & radio) and politics were predominantly discussed. This suggests that these messages could be used to extract opinions, attitudes, and sentiments towards these topics. This opens up possibilities to collect a considerable amount of information in a quick way without any response burden. The major problem in social media is discriminating the informative from the non-informative messages. Because of the large share of the non-informative 'babble' messages, use of the more serious (informative) messages is negatively affected. Text mining approaches used to automatically differentiate between both groups of messages have not been very successful so far and require further research.

Another potential use of social media messages is sentiment analysis. Access to over 1.6 billion public messages written in Dutch from a large number of social media sites was obtained through infrastructure provided by Coosto (2012). Messages are sourced from the largest social media sites including Twitter, Facebook, Hyves, Google+, and LinkedIn, but also from numerous public weblogs and forums. The overall profile of the number of messages created per day revealed that from June 2010 onwards increasing number of messages were generated in the Netherlands on a daily basis. The latter date corresponds to the onset of the increase of public Twitter messages written in Dutch. We therefore used June 2010 as the starting date for our studies with August 2012 as the end date. With a query language and a web interface, messages were selected from the database. The sentiment of each message was automatically determined 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.

Since several studies have been performed in English speaking countries attempting to link the sentiment in social media with consumer sentiment (O'Connor et al. 2010, Lansdall-Welfare et al. 2012) we were interested in studying this approach for the Netherlands. We started by testing a wide range of words of which it could be expected that they somehow correlated with consumer sentiment; such as 'buy' and 'mortgage'. This proved very difficult. Some words were hardly used and others showed no clear or stable dependence. Large differences in the amount and the sentiment of messages were observed particularly for words regularly occurring in more popular news (such as entertainment and showbiz related). This prompted us to try another approach: using very general terms. Interestingly, this general approach worked quite well. These queries returned very large amounts of messages, around 600 million for the Dutch articles and 1.2 billion for the 10 most frequently used Dutch words for the period studied, of which the overall sentiment was analysed. A British study used the term 'Mood of the nation' for this kind of general sentiment (Lansdall-Welfare et al. 2012). The monthly sentiment for the period June 2010 till August 2012 derived from Dutch social media messages was found to correlate very strongly (0.83) with the officially determined monthly Dutch consumer confidence (Statistics Netherlands 2013) and with the sentiment for the sub-indicator of the attitude towards the economic climate (0.88). Both official indicators are based on a sample survey in which 1500 people are interviewed each month. Figure 4 displays the survey-based series and the corresponding Dutch social media sentiment findings. Both series relate very well except for the month of December, where a much more positive attitude is found in social media. Removal of all messages including the (Dutch) words Christmas and references towards new year and new year's eve reduced these peaks and increased the correlation to 0.90. Clearly, the large number of positive messages created in December increased the overall positive sentiment during that period. This high correlation is remarkable, as the populations from which the data are obtained are different. Dutch consumer confidence is obtained from a random sample from the population register, while the sentiment in Dutch social media messages shown in Figure 4 is derived from around 30 million *messages* generated each month. These messages are created by a considerable part of the population, 70% according to Eurostat (2012), but not all messages are written in Dutch and different users post varying numbers of messages. Previous work by the authors revealed that the number of Twitter messages can vary from 60 per day to not even one message a month (Daas et al. 2012a).

3. Discussion

The experience gained through the case studies reveals several issues that need to be addressed when exploiting Big Data for use in official statistics. These issues are discussed below.

Data exploration

Typically Big Data sets are made available to us, rather than designed by us. Their contents and structure need to be understood prior to using the data for analysis. This is called data exploration, often involving visualisation methods (Zikopoulos et al. 2012, ch. 7). Recently some visualisation methods have emerged that are particularly suited to Big

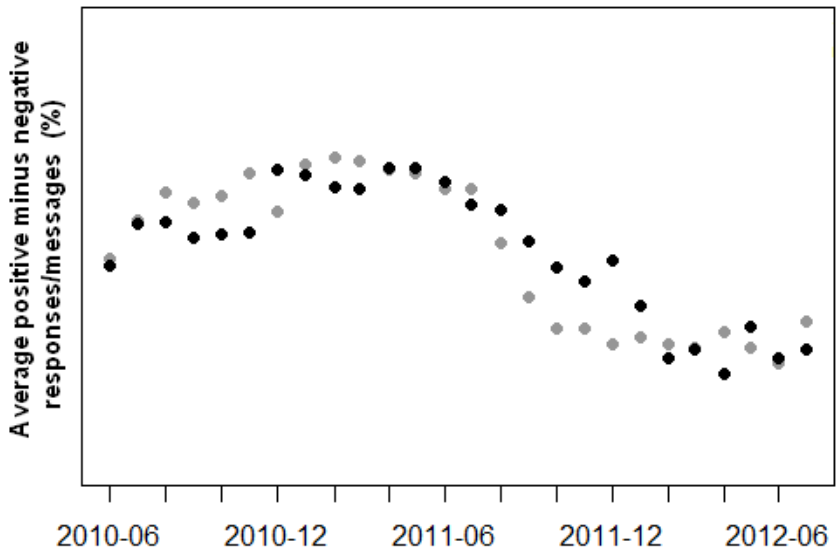

Figure 4. Dutch consumer confidence (grey) and the overall sentiment in Dutch social media messages on a monthly basis (black). Dutch articles are used as search terms. The social media sentiment in December is considerably more positive compared to the sentiment in the months before and after.

Data. Examples are tableplots (Tennekes et al. 2013) for 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 data (Daas et al. 2012b). Data exploration is aimed at revealing data structure and patterns, and assessment of quality including exposure of errors, anomalies and missing data.

Missing data

Despite the enormous amounts of data generated each day, the traffic loop detection data clearly suffers from missing data (section 2.1). By studying data on a minute-by-minute level we discovered that some data were missing. If we had analysed data aggregated at hourly or daily levels, we would have reduced the amount of data studied but would not have noticed this issue. Since our office plans to use the NDW data to produce traffic and transport statistics we need to find ways to cope with the missing data problem and simultaneously reduce the amount of data to a manageable level. Other data sources are susceptible to missing data too. In the social media example, server downtime and network outages can lead to missing data. Current efforts focus on statistical modelling able to cope with missing data, and the development of information extraction and aggregation methods.

Volatility and resolution

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 –from a statistical point of view– not very informative as they occur at too high a resolution. Similarly, sentiment analysis at a daily basis may suffer from volatility that is not seen at weekly or monthly intervals (Daas et al. 2012b; O'Connor et

al. 2010) . It is therefore recommended to develop statistical methods able to cope with volatile behaviour. Possibilities under consideration are moving averages and advanced filtering techniques (e.g. a Kalman filter or time-series modelling).

Representativity/Selectivity
The analyses in section 2 apply to traffic on roads equipped with traffic loop sensors, and to sentiment analysis of people who post Dutch messages on social media web sites. These are subpopulations of respectively all traffic on Dutch roads, and of all people in the Netherlands. The subpopulations covered by these Big Data sources are not target populations for official statistics. Therefore the data are likely to be selective, not representative of a relevant target population. Representativity of Big Data could be assessed through careful comparison of characteristics of the covered population and the target population. This may prove problematic, as often there are no characteristics readily available to conduct such comparison. For example, little is known about the people posting on social media. Often only their user name 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 addressed if necessary. This could be achieved through predictive modelling, using a wide variety of algorithms known from statistical learning and data mining. These are modelling methods traditionally not used in official statistics. Buelens et al. (2012) explore some possibilities for applications of data mining methods in official statistics.

Other issues
Long-term stability may be a problem when using Big Data. Typically, statistics for policy making and evaluation are required for extended periods of time, often covering many years. The Big Data sources encountered so far seem subject to frequent modifications, possibly compromising their long term use. Privacy and data ownership are other issues that need to be addressed, as many potential Big Data sources are collected by non-governmental organisations, a situation that may not be covered by existing legislation. Finally, dedicated and specialized computing infrastructures are required to cope with Big Data.

4. Conclusions

The official statistics community can benefit greatly from the possibilities offered by Big Data, but must invest in research and skills development. Various new areas of expertise are needed to fully exploit the information contained in Big Data. In particular, knowledge is required from the fields of register-based statistics (Bakker and Daas 2012), mining of massive data sets (Rajaraman and Ullman 2011), and the new emerging discipline commonly referred to as 'Data Science' (Loukides 2011). We expect to see some Dutch official statistics being based on Big Data in the coming years, and are working towards this aim. 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.

References

Bakker, B.F.M. and Daas, P. (2012). Methodological challenges of register-based research. Statistica Neerlandica, 66, 2-7.

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

Buelens, B., Boonstra, H.J., Van den Brakel, J., and Daas, P. (2012). Shifting paradigms in official statistics: from design-based to model-based to algorithmic inference. Discussion paper 201218, Statistics Netherlands, The Hague/Heerlen. Available at: URL=http://www.cbs.nl/NR/rdonlyres/A94F8139-3DEE-45E3-AE38-772F8869DD8C/0/201218x10pub.pdf (Accessed January 2013).

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: URL=http://www.worldbank.org/wb/Big-data-pc-2012-12-12.pdf (Accessed January 2013).

Coosto (2012). The facts webpage. Available at: URL=http://www.coosto.nl/home/about/feiten and in English at URL=http://www.coosto.co.uk/home/about/facts (Accessed January 2013.)

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

Daas, P.J.H., Roos, M., van de Ven, M., and Neroni, J. (2012a). Twitter as a potential data source for statistics. Discussion paper 201221, The Hague/Heerlen: Statistics Netherlands. Available at: URL=http://www.cbs.nl/NR/rdonlyres/04B7DD23-5443-4F98-B466-1C67AAA19527/0/201221x10pub.pdf (Accessed December 2012.)

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

Eurostat (2012). Internet access and use. Eurostat newsrelease 185/2012, December 18. Available at: URL=http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF (Accessed January 2013.)

Golder, S.A. and Macy, M.W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science, 30, 1878-1881.

Groves, R.M. (2011). Three Eras of Survey Research. Public Opinion Quarterly, 75, 861-871.

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

Loukides, M. (2010). What is Data Science? O'Reilly Radar Report, Available at: URL= http://cdn.oreilly.com/radar/2010/06/What_is_Data_Science.pdf (Accessed December 2012.)

Lynch, C. (2008). Big data: How do your data grow? Nature 455, 28-29.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. Report of the McKinsey Global Institute, McKinsey & Company. Available at: URL=http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation (Accessed December 2012.)

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

O'Connor, B., Balasubramanyan, R., Routledge, B.R., and Smith, N.A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. Carnegie Mellon University, Research Showcase. Available at= http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf (Accessed January 2013.)

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Statistics Netherlands (2013). Consumer confidence, economic climate and willingness to buy. Available at: URL= http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLEN&PA=7388eng&LA=EN (Accessed January 2013.)

Tennekes, M., de Jonge, E., and Daas, P.J.H. (2013). Visualizing and Inspecting Large Datasets with Tableplots. Journal of Data Science, 11, 43-58.

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