Big Data and Official Statistics

Piet Daas (Statistics Netherlands)

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 concept of 'Big Data'. Big Data are characterized by data sets of increasing volume, velocity and variety; the famous 3 V's. Big Data are often also largely unstructured, meaning that they have no pre-defined data model and/or do not fit well into relational tables. Apart from generating new commercial opportunities in the private sector, Big Data are also potentially very interesting as input for official statistics; either for use on their own, or in combination with more traditional data sources such as sample surveys and administrative registers. However, harvesting the information from Big Data and incorporating it into a statistical production process is not easy. They are in fact the major challenges in the area of Big Data and statistics.

1) How to turn the data in Big Data into information is the first challenge. It is the old "needle in the haystack" problem. The information required has to be extracted from the massive amounts of (unstructured) data available. This can either be done by clever data selection and transformation processes, often referred to as 'making big data small', or by large scale 'number crunching'. For the latter a state-of-the-art IT infrastructure is essential.

2) The second major challenge is making statistics from the information uncovered. This requires integrating the information into the statistical process. Since Big Data usually do not contain the unique identifiers commonly used in the normal statistical production process, this is also not an easy task. Integration at a higher (macro)level could be a way to deal with this problem.

On top of that and to my surprise, 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 has an IT-perspective, as they predominantly focus on soft- and hardware issues, and fail to address important statistical issues such as coverage, representativeness, quality, accuracy and precision. When Big Data are being used for official statistics, it is essential that these issues are considered and dealt with accordingly. Methods developed for the statistical use of administrative data sources provide a good starting point.

This document gives a short overview of the lessons learned while studying Big Data sources at Statistics Netherlands. The areas of expertise where knowledge was extracted from are very similar to those mentioned for the discipline currently referred to as 'Data Science'. Data science incorporates techniques and theories from fields such as Math, Statistics, Knowledge discovery and Machine Learning, Visualization, and High Performance Computing. The approaches discussed below are however not limited to Big Data, they can also be applied to other (large) data sources such as administrative registers.

Need for new methods

Using enormous amounts of data is not an easy task. Solely through their size alone, getting insight into Big Data and their quality is difficult. As a result of this, the data exploration phase takes considerable more time for Big Data compared to other, often more structured, large data sources. A need for 'new' exploration and analysis methods arose. The term new is placed into quotes here because many of the methods used already existed but are new in the area of official statistics. Three were found particularly fruitful, namely: Visualization methods, Text mining, and High Performance Computing.

Visualization methods, particularly tableplots, 3D heat maps and data movies, can be used to quickly gain insight into the content of a Big Data source. Tableplots display the aggregated distribution patterns of a dozen of variables in a single figure for very large data files. With this method, information on data quality and the presence and selectivity of missing data can be easily obtained. 3D heat maps have been used to study the frequency at which a combination of two variables occurs, differences being indicated by heat map colours. In the third dimension, the size of the individual values of the combination of variables is displayed. The resulting 3D 'landscape' was found to be very spiky for the data files studied, indicating the need for a more smooth (less detailed) view on the data. The latter suggests that Big Data contain way to much details to be used directly for statistical purposes. Data movies are the third visualization method employed. By consecutively combining sequences of 'static' 2D plots, movies are created which, certainly for time-related changes in large datasets, are very insightful.

Text mining is the second method needed. Since many Big Data sources, such as social media and web pages, are largely composed of text, methods capable of extracting information from texts are needed. There is a multitude of software programs able to do this, but which programs or combination of programs provide information of the highest quality is very data source and topic specific. We learned the hard way; by doing.

High Performance Computing is the third method considered very useful. Examples of this approach are parallel processing in R and analysing large amounts of unstructured data on distributed systems (in 'the cloud'). The potential of this approach was demonstrated by the large scale analysis of social media messages described in the next section. Unfortunately this method can not be applied within the secure, but rather closed, environment at our office. To increase our knowledge on this essential topic a collaboration with the High Performance Computing centre SARA in Amsterdam is initiated.

Examples of studies performed

Two examples of the statistical analysis of Big Data are discussed in this section. The Big Data sources studied are traffic loop detection data and social media messages.

In the Netherlands, about 80 million traffic loop detection records are generated a day. These data can be used as a source of information for traffic and transport statistics and -potentially- also for statistics on other economic phenomena. The data are provided at a very detailed level. More specific, for more than 20.000 detection loops on Dutch roads, the number of passing cars in various length classes is available on a minute-by-minute bases. Length classes enable the differentiation between cars and trucks. Downside of this source is that it seriously suffers from under-coverage and selectivity. The number of vehicles detected is not available for every minute and not all (important) Dutch roads have detection loops yet. At the most detailed level, that of individual loops, the number of vehicles detected demonstrates (highly) volatile behaviour, indicating the need for a more statistical approach.

Our second example concerns the around 1 million public social media messages which are produced on a daily basis in the Netherlands. These messages are available to anyone with internet access. Social media could be a potential interesting data source because people voluntarily share information, discuss topics of interest, and contact family and friends. To answer the question if social media are an interesting data source for statistics, Dutch social media messages were studied from two points of view: content and sentiment. Studies of the content of Dutch Twitter messages (the predominant public social media message in the Netherlands at the time of our study) revealed that nearly 50% of the messages were composed of 'pointless babble'. The remainder predominantly discussed spare time activities (10%), work (7%), media (TV & radio; 5%) and politics (3%). Use of these, more serious, messages was hampered by the less serious 'babble' messages. The latter also negatively affected our text mining approaches. Determination of the sentiment in social media messages revealed a very interesting potential use of this data source for statistics. The sentiment in Dutch social media messages was found to be highly correlated with Dutch consumer confidence; in particular with the sentiment towards the economic situation. The latter relation was stable on a monthly and on a weekly basis. Daily figures, however, displayed highly volatile behaviour.

In the above an overview is given of the current start-of-the-art of Big data research at Statistics Netherlands. More information can be obtained by contacting the author at: pjh.daas@cbs.nl.

Setting the Priorities for Statistical Modernisation

Steven Vale (UNECE)

The key priorities for the modernisation of statistical production are the management and implementation of relevant standards, and the development of a “plug and play” approach to producing statistics.

This was the conclusion of a workshop held in Geneva on 7-8 November, bringing together heads of statistical organisations and representatives of expert groups dealing with standards, technology and methodology.

The workshop started with a keynote presentation from Gosse van der Veen, the head of Statistics Netherlands, and chair of the High-Level Group for the Modernisation of Statistical Production and Services. He outlined the challenges facing official statistics, and a strategy to tackle them.

The workshop reviewed progress on developing a Generic Statistical Information Model (GSIM). This was identified as a key priority by a similar workshop in 2011. The GSIM was presented, and issues about communication and implementation were discussed. The GSIM will be released in December 2012, completing the set of standards needed to modernise official statistics. The implementation of these standards will be a key theme for the coming years.

Developing an architecture based on GSIM and related standards, to provide the basis for a modular approach to statistical production (“plug and play”) was also identified as a key priority.

References:

HLG: www1.unece.org/stat/platform/display/hlgbas

GSIM: www1.unece.org/stat/platform/x/SwCPAw