"Re-make/Re-model": Should big data change the modelling paradigm in official statistics?

Braaksma, Barteld and Zeelenberg, Kees

Statistics Netherlands, Statistics Netherlands

2015

Statistical Journal of the IAOS 31 (2015) 193–202
DOI 10.3233/SJI-150892
IOS Press

"Re-make/Re-model": Should big data change the modelling paradigm in official statistics?[1]

Barteld Braaksma[a,*] and Kees Zeelenberg[b]
[a]*Innovation Program, Statistics Netherlands, 2490 HA Den Haag, The Netherlands*
[b]*Methods and Statistical Policies, Statistics Netherlands, 2490 HA Den Haag, The Netherlands*

Abstract. Big data offers many opportunities for official statistics: for example increased resolution, better timeliness, and new statistical outputs. But there are also many challenges: uncontrolled changes in sources that threaten continuity, lack of identifiers that impedes linking to population frames, and data that refers only indirectly to phenomena of statistical interest. We discuss two approaches to deal with these challenges and opportunities.
First, we may accept big data for what they are: an imperfect, yet timely, indicator of phenomena in society. These data exist and that's why they are interesting. Secondly, we may extend this approach by explicit modelling. New methods like machine-learning techniques can be considered alongside more traditional methods like Bayesian techniques.
National statistical institutes have always been reluctant to use models, apart from specific cases like small-area estimates. Based on the experience at Statistics Netherlands we argue that NSIs should not be afraid to use models, provided that their use is documented and made transparent to users. Moreover, the primary purpose of an NSI is to describe society; we should refrain from making forecasts. The models used should therefore rely on actually observed data and they should be validated extensively.

Keywords: Big data, model-based statistics

1. Introduction

Big data come in high volume, high velocity and high variety; examples are web scraping, twitter messages, mobile phone call detail records, traffic-loop data, and banking transactions. This leads to opportunities for new statistics or redesign of existing statistics. Their high volume may lead to better accuracy and more details, their high velocity may lead to more frequent and timelier statistical estimates, and their high variety may give rise to statistics in new areas.

There are various challenges with the use of big data in official statistics, such as legal, technological, financial, methodological, and privacy-related ones; see e.g. [19,21,22]. This paper focuses on methodological challenges, in particular on the question how official statistics may be made from big data, and not on the other challenges.

At the same time, big data may be highly volatile and selective: the coverage of the population to which they refer, may change from day to day, leading to inexplicable jumps in time-series. And very often, the individual observations in big-data sets lack linking variables and so cannot be linked to other datasets or population frames. This severely limits the possibilities for correction of selectivity and volatility.

The use of big data in official statistics therefore requires other approaches. We discuss two such approaches.

[1] *Re-Make/Re-Model* is a song written by Bryan Ferry and performed by Roxy Music in 1972 (Wikipedia, 2015).

*Corresponding author: Barteld Braaksma, Manager Innovation Program, Statistics Netherlands, PO Box 24500, 2490 HA Den Haag, Netherlands. Tel.: +31 70 3374430; E-mail: b.braaksma@cbs.nl.

In the first place, we may accept big data just for what they are: an imperfect, yet very timely, indicator of developments in society. In a sense, this is what national statistical institutes (NSIs) often do: we collect data that have been assembled by the respondents and the reason why, and even just the fact that, they have been assembled is very much the same reason why they are interesting for society and thus for an NSI to collect. For short, we might argue: these data exist and that's why they are interesting.

Secondly, we may extend this approach in a more formal way by modelling these data explicitly. In recent years, many new methods for dealing with big data have been developed by mathematical and applied statisticians.

In Section 2 we briefly describe big data and the possible uses as well as some actual examples. In Section 3 we look at the first manner in which they may be used: as they are collected or assembled, i.e. as statistics in their own right. In Section 4 we discuss how models may be useful for creating information from big-data sources, and under what conditions NSIs may be using models for creating official statistics.

2. Big data

2.1. Source data for official statistics

Official statistics must be based on observations: often raw data that needs further processing and is honed to produce accurate, reliable, robust and timely information.

For many years, producers of official statistics have relied on their own data collection, using paper questionnaires, face-to-face and telephone interviews, or (somewhat less traditional) web surveys. This classical approach originates from the era of data scarcity, when official statistics institutes were among the few organisations that could gather data and disseminate information. A main advantage of the survey-based approach is that it gives full control over questions asked and populations studied. A big disadvantage is that it is rather costly and burdensome, for the surveying organisation and the respondents, respectively.

More recently, statistical institutes have started to use administrative (mostly government) registers as additional sources. Using such secondary sources reduces control over the available data, and the administrative population often does not exactly match the statistical one. However, these data are cheaper to obtain than conducting a survey as they are already present. In some countries, the access and use of secondary sources is regulated by law.

Big Data sources offer even less control. They typically consist of 'organic' data [10] collected by others, who have a non-statistical purpose for their data. For example, a statistical organization might want to use retail transaction data to provide prices for their Consumer Price Index statistics, while the data generator sees it as a way to track inventories and sales.

In this section we will look at some examples from the research and innovation program at Statistics Netherlands: social media messages, traffic-loop data, and mobile phone data; the text of these subsections is based on papers [4,5,13,17] from colleagues at Statistics Netherlands. These examples fall in the categories Social Networks and Internet of Things, as distinguished by the UN/ECE Task Team on Big Data [21]; we omit examples from their third category, Traditional Business systems (process-mediated data), since on the one hand some of the methodological and statistical problems for this category resemble those for administrative data and on the other hand there is not yet much experience with the more complicated types of data in this category, such as banking transactions. In particular, we discuss actual or possible uses in official statistics and some issues that arise when analysing these data sources from an official statistics perspective. Other examples, which we will not discuss here, include web scraping, scanner data, satellite images and banking transactions.

2.2. Traffic-loop data [5,17]

In the Netherlands, approximately 100 million traffic detection loop records are generated a day. More specifically, for more than 12 thousand detection loops on Dutch roads, the number of passing cars is available on a minute-by-minute basis. The data are collected and stored by the National Data Warehouse for Traffic Information (NDW) (http://www.ndw.nu/en/), a government body which provides the data free of charge to Statistics Netherlands. A considerable part of the loops discern length classes, enabling the differentiation between, e.g., cars and trucks. Their profiles clearly reveal differences in driving behaviour.

Harvesting the vast amount of data is a major challenge for statistics; but it could result in speedier and more robust traffic statistics, including more detailed information on regional levels and increased resolution in temporal patterns. This is also likely indicative of changes in economic activity in a broader sense.

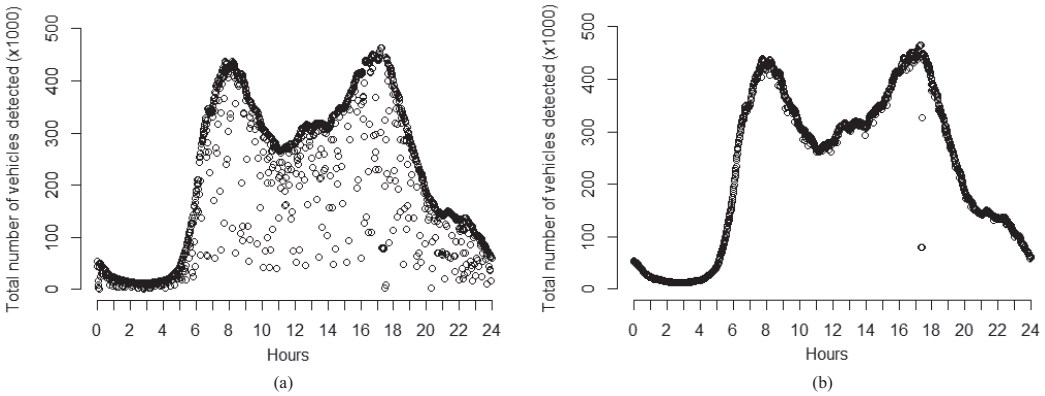

Fig. 1. Traffic distribution pattern for a single day (Thursday, 1 December 2011), aggregated over all traffic loops in five-minute blocks. Figure 1a presents raw data as recorded; Fig. 1b presents data after imputation for missing observations [5].

An issue is that this source suffers from undercoverage and selectivity. The number of vehicles detected is not available for every minute due to system failures and not all (important) Dutch roads have detection loops. Fortunately, the first can be corrected by imputing the absent data with data that is reported by the same loop during a 5-minute interval before or after that minute (see Fig. 1). Coverage is improving over time. Gradually more and more roads have detection loops, enabling a more complete coverage of the most important Dutch roads and reducing selectivity. In one year more than two thousand loops were added.

Some detection loops are linked to weigh-in-motion stations, which automatically measure the weight of the vehicle while driving and which are combined with cameras that record the license plate. One very important weigh station is in the highway connecting the port of Rotterdam to the rest of the Netherlands. In the future, these measurements may be used to estimate the weight of the transported goods. Statistical applications may then be very rapid estimates of goods transported from ports or exported and imported across land boundaries. Or they may even be used to create a rough indicator of economic activity [17].

2.3. Social media messages [4]

Social media is a data source where people voluntarily share information, discuss topics of interest, and contact family and friends. 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, but collecting them all is obviously a huge task. The social media data analysed by Statistics Netherlands were provided by the company Coosto, which routinely collects all Dutch social media messages. In addition, they provide some extra information, like assigning a sentiment score to individual messages or adding information about the place of origin of a message.

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. 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' (see Fig. 2). 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. The latter also negatively affected text mining studies.

The sentiment in Dutch social media messages was found to be highly correlated with Dutch consumer confidence [4]. Facebook gave the best overall results. The observed sentiment was stable on a monthly and weekly basis, but daily figures displayed highly volatile behaviour. Thus it might become possible to produce useful weekly sentiment indicators, even on the first working day after the week studied.

2.4. Mobile phone data [13]

Nowadays, people carry mobile phones with them everywhere and use their phones throughout the day. To manage the phone traffic, a lot of data needs to be processed by mobile phone companies. This data is very closely associated with behaviour of people; behaviour that is of interest for official statistics. For example, the phone traffic is relayed through geograph-

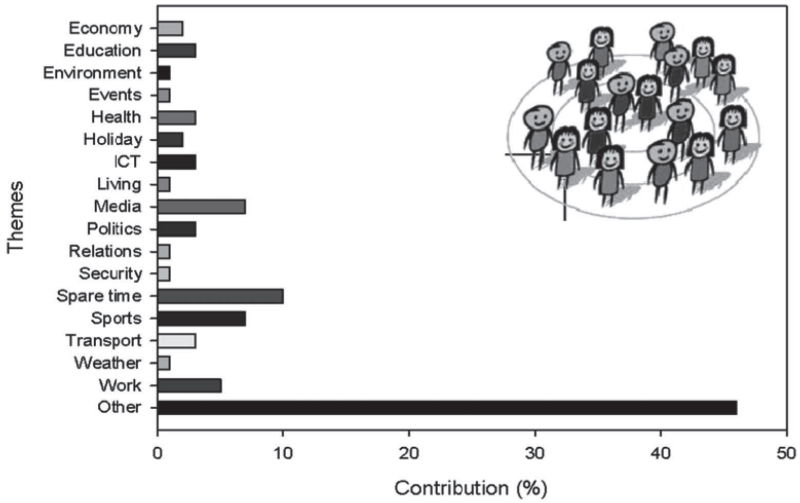

Fig. 2. Distribution of Dutch Twitter messages according to statistical theme. The themes are those identified in the annual work program of Statistics Netherlands; one extra theme, "Media" was added because of the number of tweets relating to this topic. The "Other" category refers to tweets that could not be related to any theme [4].

ically distributed phone masts, which enables determination of the location of phone users. The relaying mast, however, may change several times during a call: nontrivial location algorithms are needed.

Several uses for official statistics may be envisaged, including inbound tourism [11] and daytime population [20]. 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.

3. Big data as statistics

Big data lead to new opportunities for new statistics or redesign of existing statistics. Their high volume may lead to better accuracy and more details, their high velocity may lead to more frequent and timelier statistical estimates, and their high variety may give opportunities for statistics in new areas.

At the same time, big data may be highly volatile and selective: the coverage of the population to which they refer may change from day to day, leading to inexplicable jumps in time-series. And very often, the individual observations in these big-data sets lack linking variables and so cannot be linked to other datasets or population frames. This severely limits the possibilities for correction of selectivity and volatility using traditional methods.

For example, phone calls usually relate to persons, but how to interpret their signals is far from obvious. People may carry multiple phones or none, children use phones registered to their parents, phones may be switched off, etcetera. Moreover, the way people use their phones may change over time, depending on changes in billing, technical advances, and preferences for alternative communication tools, among other things. For social media messages, similar issues may arise when trying to identify characteristics of their authors.

Many Big Data sources are composed of event-driven observational data which are not designed for data analysis. They lack well-defined target populations, data structures and quality guarantees. This makes it hard to apply traditional statistical methods, based on sampling theory.

In this section we discuss one way how NSIs may deal with these statistical problems, namely whether we might regard big-data aggregates as statistics in their own right. We may accept the big data just for what they are: an imperfect, yet very timely, indicator of developments in society. In a general sense, this is what NSIs often do: we collect data that have been assembled by the respondents and the reason why, and even just the fact that they have been assembled, is very much the same reason why they are interesting for society and thus for an NSI to collect. For short, we might argue: these data exist and that's why they are interesting.

This is perhaps most obvious with social media messages, and indicators derived from them. Opinions expressed on Twitter or Facebook already play a role, and

sometimes an important role, in public debates. For example, the website (http://www.nos.nl/) of the Dutch radio and television system often adds twitter messages sent by the public to its news items, and so these twitter messages become part of the news and of public discussion.

Also the sentiment indicator based on social media messages, discussed in the previous section, is an example. It has been shown that this indicator is highly correlated with more traditional estimates of consumer confidence. Therefore we may conclude that this indicator is relevant. However, the social media-based sentiment indicator does not track exactly the traditional indicator. On the other hand, the traditional way of making consumer-confidence statistics is by means of a telephone survey, and these statistics contain therefore sampling errors, and, perhaps worse, also non-sampling errors. The important point here is that the traditional consumer-confidence indicator is not an exact measure of consumer confidence, because of sampling errors, and possibly even has a bias, because of non-sampling errors. Thus, it would be more appropriate to say that the social media sentiment indicator and the traditional indicator both are estimates of 'the mood of the nation', and we should not consider one of these to be the exact and undisputable truth.

One should not forget that apart from accuracy, quality has other aspects: relevance, timeliness, accessibility, comparability and coherence [6, 7]. Since the social media indicator clearly can be produced much more frequently and timely, it scores higher on the aspect of timeliness. On the other hand, comparability may be much harder to maintain, since participation in social media may change or even show large fluctuations over time; and methods similar to non-response correction methods in surveys, may have to be used to correct for this. Still, even if the social-media sentiment indicator might score lower on relevance or accuracy, it may be because of its timeliness still be useful for society if an NSI produces it as an official statistic.

The other examples of big data presented in Section 2 can also be judged according to the usual quality dimensions.

For example as described in Subsection 2.2, traffic-loop data may be used to produce very rapid estimates of traffic intensity and possibly also of the quantity of goods transported, exported and imported. Since quantities will be based on the weight of the transported goods, the bias component of its accuracy may be higher than that of the traditional estimate derived from a survey among transport companies, but because its coverage will be nearly complete, the variance component will be nearly zero. And such a very rapid estimate may be highly relevant.

With mobile-phone data, there may be more problems of representativeness: some persons carry more than one mobile phone, some phones may be switched off, and background characteristics are not known or imperfect because of prepaid phones, company phones, and children's phones registered to parents. There can also be accuracy issues when mapping phone masts to statistically relevant geographical areas: often they do not overlap perfectly. This problem becomes more pronounced when going to higher levels of detail, where to some extent model-based decisions need to be made for assigning phone calls to areas.

4. Official statistics from models for big data

In this section we discuss how models may be useful for creating information from big-data sources, and under what conditions NSIs may be using models for creating official statistics.

4.1. Design-based, model-assisted and model-based methods

We follow the well-known distinction between *design-based methods*, *model-assisted methods* and *model-based methods*. Design-based methods are the methods that strictly conform to a survey model, where respondents are sampled according to known probabilities, and the statistician uses these probabilities to compute an unbiased estimator of some population characteristics, such as average income. Model-assisted methods use a model that captures some prior information about the population to increase the precision of the estimates; however, if the model is incorrect, then the estimates are still unbiased when taking only the design into account. The examples of big data given in Section 3 rely mostly on the data as collected supplemented with obvious corrections for probabilities of observation, and thus fall in the categories of design-based or model-assisted methods.

Model-based methods, however, rely on the correctness of the model: the estimates are biased if the model does not hold. As an example, suppose we want to estimate consumer confidence in a certain period, and that we have a traditional survey sample for which consumer confidence according to the correct statistical concept is observed, but also a social media source

where a sentiment score can be attached to individual messages by applying a certain algorithm. A model-assisted approach would be to use the social media source data as auxiliary variables in a regression estimator. Even if the model that relates consumer confidence to sentiment scores does not hold perfectly, the resulting estimator is still approximately unbiased under the sampling design. A simple example of a model-based estimator would be to aggregate all the individual sentiment scores in the social media source, and use this as an estimate for consumer confidence. The implicit model here is that sentiment in the social media source is equal to consumer confidence in the statistical sense. If this model does not hold, then the resulting estimate will be biased. Of course, if we actually do have both types of data, the sample and the social media data, it would not be efficient to use only the latter data in a model-based estimator. But it may be much cheaper to not sample at all and to use only the big data source. The response burden on persons in the sample may also be a barrier to maintain a survey if a suitable alternative is available.

National statistical institutes have always been reluctant to use model-based methods in official statistics. They have relied on censuses and surveys, using mostly design-based and model-assisted methods. Yet, in specific statistical areas, NSIs have used model-based methods, e.g. in making small-area estimates, in correcting for non-response and selectivity, in computing seasonally-adjusted time series, and in making preliminary macro-economic estimates. And, in fact, common techniques like imputation of missing data often rely on some model assumptions. So in a sense, models are already being used in official statistics. But very often, these models remain implicit and are not being emphasized in the documentation and the dissemination. Therefore, in general NSIs should not be scared to use model-based methods for treating big-data sources. In the next subsections we will look at how this might be done.

4.2. Coverage and selectivity

Big data may be highly volatile and selective: the coverage of the population to which they refer may change from day to day, leading to inexplicable jumps in time-series. And very often, the individual observations in these big-data sets lack linking variables and so cannot be linked to other datasets or population frames. This severely limits the possibilities for correction of selectivity and volatility. On the other hand, for many phenomena where we have big data, we also have other information, such as survey data for a small part of the population, and prior information from other sources.

One way to go then is to use big data together with such additional information and see whether we can model the phenomenon that we want to describe. In recent years there has been a surge in mathematical statistics in developing advanced new methods for big data. They come in various flavours, such as high-dimensional regression, machine-learning techniques, graphical modelling, data science, and Bayesian networks [1,3,8,15,23]. Also, more traditional methods, such as Bayesian techniques, filtering algorithms and multi-level (hierarchical) models have appeared to be useful [9].

Another strategy is to take inspiration from the way National accounts are commonly compiled. Many sources which are in themselves incomplete, imperfect and/or partly overlapping are integrated, using a conceptual reference frame to obtain a comprehensive picture of the whole economy, while applying many checks and balances. In the same way, big data and other sources that in themselves are incomplete or biased may be combined together to yield a complete and unbiased picture pertaining to a certain phenomenon.

More generally, one might say that big data are a case where we have insufficient information about the relations of the data source to the statistical phenomena we want to describe. This is often caused by lack of information about the data-generating process itself. Models are then useful to formulate explicit assumptions about these relations, and to estimate selectivity or coverage issues. For example, one way to reduce possible selectivity in a social media source could be to profile individual accounts in order to find out more about background characteristics. If we can determine whether an account belongs to a man or a woman, we should be able to better deal with gender bias in sentiment. Techniques to do this have already been developed and are becoming increasingly sophisticated. The same applies to age distribution, education level and geographical location. Coverage issues with individual social media sources can be reduced by combining multiple sources; and the sensible way to do this is through using a model, for example a multiple regression model or a logit model if we have information about the composition of the various sources. Another example is the use of a Bayesian filter to reduce volatility, as presented below in Section 4.4.

4.3. Quality, objectivity and reliability

NSIs must, as producers of official statistics, be careful in the application of model-based methods. The public should not have to worry about the quality of official statistics, as formulated in the mission statement of the European Statistical System:

> *"We provide the European Union, the world and the public with independent high quality information on the economy and society on European, national and regional levels and make the information available to everyone for decision-making purposes, research and debate."*

Objectivity and *Reliability* are among the principles of official statistics in the European Statistical Law [6] "...meaning that statistics must be developed, produced and disseminated in a systematic, reliable and unbiased manner." And the European Statistics Code of Practice [7] says: "European Statistics accurately and reliably portray reality." Other international declarations, such as those of the ISI [12] and the UN [18], but also national statistical laws such as those of the Netherlands, have similar principles.

When using models, we can interpret these two principles as follows. The principle of objectivity means that the data that are being used to estimate the model should refer to the phenomenon that one is describing; in other words, the objects and the populations for the model correspond to the statistical phenomenon at hand. Data from the past may be used to estimate the model, but official statistical estimates based on the model never go beyond the present time period; so for an NSI now-casting is allowed, but not forecasting and policy analyses. Of course this is different for a forecasting agency or a policy-evaluation agency, whose purpose is exactly to go beyond the present period or present context. We believe that even if official statistics and policy evaluation is combined, for example in one report or even as is the case with some NSIs in one organization, it is always desirable to distinguish official statistics, which describe what has actually happened, from policy evaluation which deals with "what-if" situations.

The principle of reliability means that we must prevent having to revise official statistical data just because the model changes, e.g. because it breaks down (*model failure*). In particular for time-series models we must be on guard, because model failure may lead to an incorrect identification of turning points in the series.

Also we should refrain from using behavioural models, because these are prone to model failure: it is almost certain that at some time in the future, any behavioural model will fail because behaviour of economic and social agents has changed. An additional reason to avoid behavioural models is that we must prevent situations where an external researcher finds good results when fitting a certain model, but, unknowingly to the researcher, that same model had been used by the NSI to create the very data that have been used by the external researcher. Again, this is different for a forecasting agency or a policy-evaluation agency.

The principles of objectivity and reliability also lead to some methodological principles for model-based methods. In particular, model building should be accompanied by extensive specification tests, in order to ensure that the model is robust.

Based on these principles, Statistics Netherlands has developed guidelines [2] for the use of models in the production of official statistics. Many, if not most, examples in official statistics where models have been used, conform to these guidelines. So, despite the above warnings, we believe that there is room for using models also in the production of official statistics from big data.

4.4. Examples

Below we present a few examples of model-based approaches using big data. Note that all of these examples are still in the research phase. The authors of this paper are not aware of cases where similar approaches are already used in regular production of official statistics.

4.4.1. Analysis of individual traffic loops

At the level of individual loops, the number of detected vehicles displays highly volatile behaviour. This is largely due to the unpredictability of traffic at the level of individual vehicles. Sophisticated techniques are needed to identify patterns and produce meaningful statistics. One approach taken by researchers at Statistics Netherlands was to consider Bayesian recursive filters, assuming the underlying raw traffic loop data obeys a Poisson distribution (see Fig. 3).

4.4.2. Traffic loops data and regional economic activity

Does traffic intensity contain relevant information on regional economic activity? This is an interesting question, which was tested using traffic loop data in the region of Eindhoven, an important manufacturing area in the Netherlands [17]. Data from the manufacturing

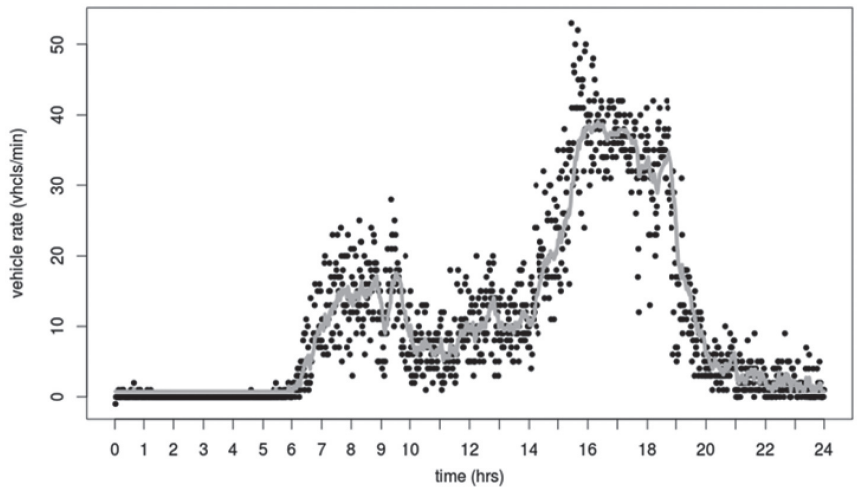

Fig. 3. Result (grey line) of application of a Bayesian recursive filter to raw data (black dots) from a single traffic detection loop, assuming that they obey a Poisson distribution [5].

sentiment survey was used as a benchmark, since this is known to be a very good business cycle indicator, with a strong and proven relation to short-term economic developments. The survey outcomes are available per province, and Eindhoven is the dominant region in the province of North Brabant. This means that the data from this survey should have a strong connection to economic activity in the Eindhoven region.

The analysis was done using three different techniques: a straightforward data selection and aggregation process, an Independent Component Analysis (ICA) algorithm and an Empirical Mode Decomposition (EMD) algorithm. All three techniques yielded similar results, but the latter (EMD) appeared to show the best overall performance (see Fig. 4).

The evolution of the traffic intensity indicator tracks that of expected production development amazingly well. Peaks and troughs coincide, meaning that the traffic intensity index should be able to signal important turning points in economic activity.

With some further processing, notably seasonal adjustment, the coherence between the two series can probably be improved even further. Another important option is to perform trend-cycle decomposition, which could improve focus on the business cycle component and remove some noise. Unfortunately the traffic intensity series is too short at the moment for both types of filtering.

4.4.3. Google Trends for nowcasting

In [3], the authors show how to use search engine data from Google Trends to 'predict the present', also known as nowcasting. They present various examples of economic indicators including automobile sales, unemployment claims, travel destination planning, and consumer confidence.

In most cases, they apply simple autoregressive models incorporating appropriate Google Trends search terms as predictors. For nowcasting consumer confidence they use a Bayesian regression model, since in that case it is not so clear in advance which search terms to use.

They found that already their simple models that include relevant Google Trends variables tend to outperform models that exclude these predictors by 5% to 20%. No claims to perfection or exhaustiveness are made, but these preliminary results indicate that it is worthwhile to pursue this model-based path further.

On the other hand, we should be cautious with interpreting search-term based results. A couple of years ago there was a lot of enthusiasm concerning Google Flu, but more recently the nowcasting performance of Google Flu has decreased significantly [14]. Google have also been criticized for not being transparent: they have not revealed the search terms used in Google Flu, which inhibits a sound scientific debate and cross-validation by peers.

In fact, this last point has more general significance. One of the items in the European Code of Practice [7], is that NSIs should warn the public and policy makers when statistical results are being used inappropriately or are being misrepresented. As emphasized in [8,16], with big data it is easy to find spurious results, and there is a role for NSIs as *statistical authorities* to offer best practices for analysing big data.

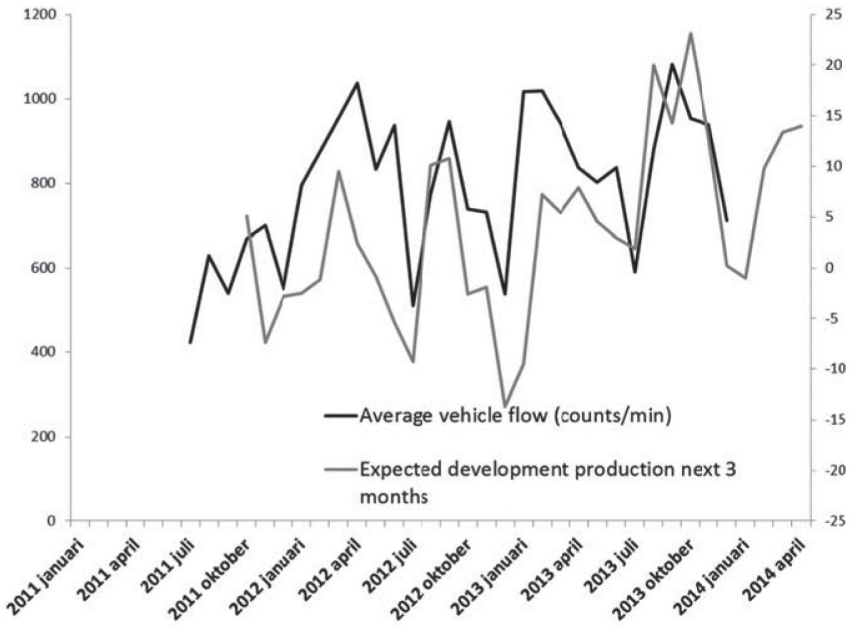

Fig. 4. EMD-filtered monthly indicator of average rush hour vehicle flow in the Eindhoven area. compared to expected production development in the manufacturing industry for the province of North Brabant. Computed correlation is 0.523 [17].

5. Conclusion

There are three main conclusions.

First, big data come in high volume, high velocity and high variety. This leads to new opportunities for new statistics or redesign of existing statistics:

- Their high volume may lead to better accuracy and more details,
- Their high velocity may lead to more frequent and more timely statistical estimates,
- Their high variety may give opportunities for statistics in new areas.

Secondly, at least in some cases, statistics based on big data are useful in their own right, for example because they are being used in policy making or play a role in public discussion.

Thirdly, in general NSIs should not be scared to use models in producing official statistics, as they have apparently done this before, provided these models and methods are adequately documented. So we should look more closely at how models may be used to produce official statistics from big data. In particular Bayesian methods and multilevel models seem promising.

On the other hand, the use of models should be made explicit. It should be documented and made transparent to our users. Also, models are not to be used indiscriminately: we should not forget that the primary purpose of an NSI is to describe, and not to prescribe or judge. So we should refrain from making forecasts and from using purely behavioural models. Also, we should be careful to avoid model failure when the assumptions underlying it break down. Therefore any model should rely on actually observed data for the period under consideration, which relates to the economic and social phenomena we are trying to describe by statistical estimates; and model building should be accompanied by extensive specification tests.

References

[1] A. Belloni, V. Chernozhukov and C. Hansen, High-dimensional methods and inference on structural and treatment effects, *Journal of Economic Perspectives* **28**(2) (2014), 29–50, doi: 10.1257/jep.28.2.29.

[2] B. Buelens, P.-P. de Wolf and K. Zeelenberg, Model-based estimation at Statistics Netherlands. Discussion Paper, Statistics Netherlands, The Hague, 2015.

[3] H. Choi and H.R. Varian, Predicting the present with Google trends, http://people.ischool.berkeley.edu/∼hal/Papers/2011/ptp.pdf, 2011.

[4] P.J.H. Daas and M.J.H. Puts, Social media sentiment and consumer confidence, Paper presented at the Workshop on using Big Data for Forecasting and Statistics, Frankfurt, 2014. https://www.ecb.europa.eu/events/pdf/conferences/140407/Daas_Puts_Sociale_media_cons_conf_Stat_Neth.pdf?409d61b733fc259971ee5beec7cedc61.

[5] P.J.H. Daas, M.J. Puts, B. Buelens and P.A.M. van den Hurk, Big Data and Official Statistics. Paper presented at

the Conference on New Techniques and Technologies for Statistics, 5–7 March 2013, Brussels. http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf, 2013.

[6] European Union, Regulation on European statistics, Official Journal of the European Union, L 87 (31 March 2009), 164–173, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32009R0223:EN:NOT, 2009.

[7] European Union, Code of Practice for European Statistics, revised edition, Eurostat, Luxembourg. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice, 2005/2011.

[8] J. Fan, F. Han and H. Liu, Challenges of big data analysis, *National Science Review* **1**(2), 293–314, doi: 10.1093/nsr/nwt032, 2014.

[9] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari and D.B. Rubin, Bayesian Data Analysis, 3e, Chapman and Hall/CRC, 2013.

[10] R.M. Groves, Three eras of survey research, *Public Opinion Quarterly* **75** (2011), 861–871, doi: 10.1093/poq/nfr057.

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

[12] International Statistical Institute, Declaration on Professional Ethics, revised edition, http://www.isi-web.org/about-isi/professional-ethics, 1985/2010.

[13] E. de Jonge, M. van Pelt and M. Roos, Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. Discussion paper 2012/14, Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/010F11EC-AF2F-4138-8201-2583D461D2B6/0/201214x10pub.pdf, 2012.

[14] D. Lazer, R. Kennedy, G. King and A. Vespignani, The parable of Google flu: traps in big data analysis, *Science* **343**(14) (2014), 1203–1205, doi: 10.1126/science.1248506.

[15] D.W. Nickerson and T. Rogers, Political campaigns and big data, *Journal of Economic Perspectives* **28**(2) (2014), 51–74, doi: 10.1257/jep.28.2.51.

[16] C. Reimsbach-Kounatze, The proliferation of "big data" and implications for official statistics and statistical agencies: A preliminary analysis, OECD Digital Economy Papers, No. 245, OECD Publishing, Paris. doi:10.1787/5js7t9wqzvg8-en, 2015.

[17] F.J. van Ruth, Traffic intensity as indicator of regional economic activity, Discussion paper 2014/21, Statistics Netherlands, 2014.

[18] Statistical Commission of the United Nations, Fundamental Principles of Official Statistics. http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx, 1991/2014.

[19] P. Struijs, B. Braaksma and P.J.H. Daas, Official statistics and Big Data. *Big Data & Society*, April–June 2014, pp. 1–6, doi: 10.1177/2053951714538417.

[20] M. Tennekes and M.P.W. Offermans, Daytime population estimations based on mobile phone metadata. Paper prepared for the Joint Statistical Meetings, Boston, 2014. http://www.amstat.org/meetings/jsm/2014/onlineprogram/AbstractDetails.cfm?abstractid=311959.

[21] UN-ECE High-Level Group for the Modernisation of Statistical Production and Services, What does "big data" mean for official statistics? http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622, 2013.

[22] C. Vaccari, Big Data in Official Statistics. PhD thesis, University of Camerino, 2014.

[23] H.R. Varian, Big data: New tricks for econometrics, *Journal of Economic Perspectives* **28**(2) (2014), 3–28, doi:10.1257/jep.28.2.3.