Discussion paper

Big data in official statistics

Barteld Braaksma
Kees Zeelenberg

January 2020

Content

Summary

In this paper, we describe and discuss opportunities for big data in official statistics. Big data come in high volume, high velocity and high variety. Their high volume may lead to better accuracy and more details, their high velocity may lead to more frequent and more timely statistical estimates, and their high variety may give opportunities for statistics in new areas. But there are also many challenges: there are uncontrolled changes in sources that threaten continuity and comparability, and data that refer only indirectly to phenomena of statistical interest. Furthermore, 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 variables that allow them to be linked to other datasets or population frames. This severely limits the possibilities for correction of selectivity and volatility. Also, with the advance of big data and open data, there is much more scope for disclosure of individual data, and this poses new problems for statistical institutes. So, big data may be regarded as so-called nonprobability samples. The use of such sources in official statistics requires other approaches than the traditional one based on surveys and censuses. A first approach is to accept the 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. In short, we might argue: these data exist and that's why they are interesting. A second approach is to use formal models and extract information from these data. In recent years, many new methods for dealing with big data have been developed by mathematical and applied statisticians. 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 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. On the other hand, in official statistics, models should not be used for all kinds of purposes.

Keywords

big data, model-based statistics, official statistics

1. Introduction

Big data come in high volume, high velocity and high variety; examples are web scraping, Twitter and Facebook messages, mobile-phone records, traffic-loop data, and banking transactions. This leads to opportunities for new statistics or redesign of existing statistics. The potential of big data for official statistics lies in the immense amount of information contained. For example, the sheer size might make it possible to add more details to official statistics. Also, big data sources may cover areas of society for which official statistics do not yet exist. 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., Struijs, Braaksma, and Daas(2014), UN-ECE (2013), and Vaccari (2016). This paper focuses on methodological challenges, in particular on the question of how official statistics may be produced from big data. Specifically, we look at the question: what are the best strategies for using big data in official statistics?

From a methodological perspective, big data also poses many challenges. 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.

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.

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.

The structure of this chapter is as follows. In section 2, we briefly describe big data sources and possible uses. In section 3, we look at statistics from big data as they are collected or assembled, i.e., as statistics in their own right; we describe several examples of big data as official statistics. 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; we also describe several examples of statistics derived through modelling big data. We look at how to create representative estimates and how to make the most of big data when this is difficult or impossible. We show how big data may be useful in solving several of the major challenges to official statistics, in particular the quality of national accounts (rate of growth of gross national product or GNP), the timeliness of official statistics, and the statistical analysis of complex and coherent

phenomena. We also look at disclosure control of official statistical data now that open data and big data have become so widely available.

2. Big data and official statistics

There are nowadays three types of sources for official statistics:

- Survey data: data collected by interviewing persons, households and enterprises,
- Administrative data: data from official registrations, such as the population register and the tax registrations,
- Big data: data generated by digital activities, e.g., activities on the internet, communications between devices or between devices and their operators (*internet of things*).

Traditionally, producers of official statistics have relied on their own data collection, using paper questionnaires, face-to-face and telephone interviews, or, in recent years, web surveys. The classical approach originates from the era of data scarcity, when official statistical institutes were among the few organizations that were able to 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 costly and burdensome, both for the surveying organization and the respondents.

More recently, statistical institutes have started to use administrative registers, usually assembled by government agencies as sources for official statistics. This reduces control over the data and leads to overcoverage or undercoverage of the population as well as to errors. For example, the official population register contains no data about homeless persons or illegal persons; additionally, people may not actually live at their registered address, so that the administrative population often does not exactly match the statistical one. However, these data are cheaper to obtain than survey data. In some countries, the access and use of secondary sources has even been regulated by law, so that statistical institutes have easy and free access.

Big data sources offer even less control. They typically consist of *organic* data (Groves, 2011) 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 original purpose of the data collector, e.g., a supermarket chain, is to track inventories and sales.

3. Examples of big data in official statistics

In this section, we will look at three examples of big data in official statistics: social media messages, traffic-loop data, and mobile phone data. Other examples, which we will not discuss here, include web scraping, scanner data, satellite images and banking transactions.

3.1 Traffic-loop Data (Van Ruth, 2014; Daas et al, 2013)

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 agency which provides the data to Statistics Netherlands. The detection loops discern length classes, enabling the differentiation between, e.g., cars and trucks. Their profiles clearly reveal differences in driving behavior.
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.
Unfortunately, at present this source suffers from under-coverage and selectivity. The number of vehicles detected is not available for every minute due to system failures and many Dutch roads, including some important ones, lack detection loops. Fortunately, the first problem can be corrected by smoothing, e.g., by imputing the absent data with data that is reported by the same loop during a 5-minute interval before or after that minute. And coverage is improving over time, as 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 (Van Ruth, 2017).

3.2 Social Media Messages (Daas and Puts, 2014)

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 analyzed by Statistics Netherlands were provided by the company Coosto (https://www.coosto.com/en), 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 a useful data source for statistics, Dutch social media messages were studied from two perspectives: content and sentiment. Studies of the content of Dutch Twitter messages, which was the predominant public social media message in the Netherlands at the time, revealed that nearly 50% of those messages were composed of "pointless babble." The remainder predominantly discussed spare time activities (10%), work (7%), media (5%) and politics (3%). Use of these, more serious, messages was hampered by the less serious "babble" messages. 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 (Daas and Puts, 2014). Facebook gave the best overall results. The observed sentiment was stable on a monthly and weekly basis, but daily figures displayed highly volatile behavior. Thus, it might become possible to produce useful weekly sentiment indicators, even on the first working day after the week studied.

3.3 Mobile Phone Data (De Jonge, Van Pelt, & Roos, 2012)

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 behavior of people; behavior that is of interest for official statistics. For example, the phone traffic is relayed through geographically distributed phone masts, which enables determination of the location of phone users. The relaying mast, however, may change several times during a call: nontrivial location algorithms are needed. Several uses for official statistics may be envisaged, including inbound tourism (Heerschap et al, 2014) and daytime population (Tennekes and Offermans, 2014). 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. Obviously, we have to take into account several issues when considering statistical uses of mobile phone data. For example, the group of mobile phone users is selective when compared with the population: young children do not usually carry a mobile phone, approximately 40 percent of the people over 65 do not have a mobile phone, whereas more than 95 percent of the young people (12 – 30 years) have a mobile phone (Telecompaper, 2015). Also, we may not have the data from all mobile-phone providers, which will create additional selectivity. This means that we have to be careful when interpreting the data or when developing processing methods.

4. A framework for assessing the quality of big data statistics

The framework we use for evaluating big data is, following the EU Statistical Law, that of statistical quality (relevance, accuracy, timeliness, accessibility, comparability, and coherence) and statistical principles (independence, impartiality, objectivity, reliability, confidentiality, and cost effectiveness). In particular, we focus on

- *accuracy, objectivity and reliability*, since these are fundamental for official statistics: if statistics do not describe society accurately enough or are not objective or reliable, they are essentially useless;
- *relevance, timeliness, accessibility, comparability and coherence*.

In the discussion below we distinguish between survey data (data based on a probability sample from a well-defined population), census data (data based on a complete enumeration of a population), administrative data (data from official sources, for a well-defined population), and big data (large datasets from other sources than the other three categories).

4.1 Accuracy

The accuracy of any statistic is measured by its variance and its bias. To judge these, we have to know the process by which the data have been generated. For surveys based on probability sampling, the bias is approximately zero, but the variance is positive, whereas for censuses (complete enumerations of the population), both are approximately zero. For data not based on surveys or censuses, the variance of statistics may be small, even approximately zero, if the dataset is sufficiently large, but the bias may be large, even if the size of the dataset is very large. For example, for a dataset of 150 million persons, the bias may be so large that its accuracy is the same as that of a probability sample of 500 persons. In a certain sense, there is a paradox here: the bigger the data, the bigger the chance that the statistical results are wrong. So, without further information, big data are useless, since bias cannot be ascertained, but may be large.

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. These “fuzzy” big data are often collected through some intermediary (“aggregator”) such as Google, Facebook and Coosto. They lack well-defined target populations, data structures and quality guarantees. This makes it hard to apply traditional statistical methods that are based on sampling theory. In fact, statistical applications of these big data, such as the Google Flu index and the Billion Prices Project, always refer to, and will always have to refer to, official statistical series to establish their validity.
It is therefore necessary to use additional information, either from the big data source itself, or from other data sources. The additional information must be sufficient for correction of the bias, not completely of course but such that the statistic will be relevant to users. This information may consist of characteristics of the individuals in the dataset that make post-stratification possible. However, if some important part of the population that is important is completely missing, correction will be impossible.

4.2 Models in Official Statistics

Use of additional information, requires the use of models that specify the relations between the statistics of interest and the additional information.
We follow here 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 in official statistics given above, 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.

Based on these principles, Statistics Netherlands has developed guidelines (Buelens, De Wolf and Zeelenberg, 2017) 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.

4.3 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 (EU, 2009) "... meaning that statistics must be developed, produced and disseminated in a systematic, reliable and unbiased manner". And the European Statistics Code of Practice (EU, 2005) says: "European Statistics accurately and reliably portray reality." Other international declarations, such as those of the ISI (1985) and the UN (1991), 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 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 are 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 behavioral models, because these are prone to model failure: it is almost certain that at some time in the future, any behavioral model will fail because behavior of economic and social agents has changed. An additional reason to avoid behavioral models is that we must prevent situations where an external researcher finds good results when fitting a certain model, but, unknown 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.
In addition, models are not to be used indiscriminately: we have to remember 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 making purely behavioral 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.
The principles of objectivity and reliability 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. And any use of models must be made explicit: it should be documented and made transparent to users.

4.4 Relevance

Even if we know that a big data source covers the whole population, the statistical information that can be extracted, may be limited. For example, with traffic sensors one can observe that a car is passing, but we don’t know who is in the car, who owns the car, and why he or she is driving at that spot. So, these data can be used for statistics on traffic intensity, but the relevance and the coherence with other statistics remain limited.

5. Big data as statistics

One way to implement big data in official statistics is to 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.

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 public radio and television system often adds Twitter messages from 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, as the traditional way of making consumer-confidence statistics is by means of a telephone survey, these statistics can contain sampling errors, and also, perhaps worse, 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.

Further, in addition to accuracy, quality has other aspects: relevance, timeliness, accessibility, comparability and coherence (EU, 2005, 2009). Since the social media indicator clearly can be produced much more frequently, 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 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 above can also be judged according to the usual quality dimensions.

For example, 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 lower 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.

6. Integration of Big data with other statistical sources

6.1 Big data as auxiliary data

Big data may be used as auxiliary data for statistics; in this approach, they are combined with other statistical data from surveys or from administrative sources. We may distinguish two cases:

- The big data source may be linked to the statistical data at the individual level. This is very much like the situation described in the previous section, and in general there are limited possibilities for linking.
- The big data may be linked to the statistical data at some intermediate level of aggregation, for example at the level of regions or industries. This looks very much like integration of statistical sources as in the National Accounts. For example, if a series derived from big data appears to be co-integrated with a statistical series, we may use the big data series to add more details to a statistical series or to increase its frequency or its timeliness; and thereby we increase its relevance and its timeliness. This is potentially a very promising route. It may also have an important impact on the timeliness of the National Accounts.

6.2 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.
In other words, with big data there is often 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.

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 (Belloni, Chernozhukov, & Hansen, 2014; Choi & Varian, 2011; Gelman et al, 2013; Nickerson & Rogers, 2014; Varian, 2014). Also, more traditional methods, such as Bayesian techniques, filtering algorithms and multi-level (hierarchical) models have appeared to be useful (Gelman et al, 2013).

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.

6.3 Disclosure control with big data

Traditionally, statistical disclosure control by national statistical institutes (NSIs) has focused on tables and microdata collected and produced by the NSI itself. However, big data have made it possible for private data companies to assemble databases with very detailed information on individuals and enterprises, with data coming from many sources, and linked sometimes deterministically but often also probabilistically.

This abundance of data poses new problems for statistical disclosure control, both strategic and ethical problems as well as methodological problems, which need to be addressed but for which there are at present no clear-cut solutions. For example, should NSIs, when protecting tables or data files against disclosure, take into account the possibility that government agencies and private companies will use the NSI data to enrich their own databases and so get to know more about their citizens or customers? And what if these enriched data are used to profile citizens so that they may come under suspicion or are denied access to certain services such as loans? Should we use another methodological paradigm than we have used so far? Should we consider different disclosure scenarios? Does the

changing attitude towards privacy influence the way we should treat our published data? How much existing as well as future data should we take into account when assessing disclosure risks?
As mentioned in the Introduction, big data are often characterized by the three V's: volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Each V poses several questions and issues related to statistical disclosure control (SDC):

6.3.1 Volume

How will NSIs deal with huge amounts of data? In general they will continue to publish aggregated information. In that case the current SDC techniques might still be applicable. However, when big data (or excerpts from them) are released as microdata, the current SDC techniques no longer apply: identity disclosure is almost certain. Then methods producing synthetic data may be preferable: use the big data sets to estimate a model and use that model to generate a synthetic dataset that resembles the original big data. Or other techniques that mask the true values of sensitive variables: just create enough uncertainty about the exact values to make it less sensitive.

6.3.2 Velocity

When data become available more quickly, the processing of the data also needs to be done in less time. The current SDC techniques can be time consuming when the underlying datasets increase in size and number. Streaming data might lead to streaming statistics, but then we should be able to protect those statistics in real time.

6.3.3 Variety

With big data we might have unstructured data, distributed over different places, indirect observations (events instead of units), unclear underlying population, selectivity, etc. Current SDC methods rely on masking characteristics of individuals so that it seems as if there are more than one similar unit in the population. For example, information about the characteristic zip code might be replaced by information about the region of residence, so that there will be many more individuals with that characteristic, and identification of individuals in the dataset becomes much more difficult. But when the underlying population is not known, this is not a valid option. Uncertainty should then be attained by introducing uncertainty on the sensitive information directly, for example by rounding.

6.4 Examples of model-based and integrated approaches using big data

6.4.1 Size of the internet economy[1]

The internet is clearly an important technology that acts as both a driver and a means for changes in the economy. There are four main ways in which the internet influences the economy:

[1] This subsection in based on Oostrom et al (2016).

- Means of communication: online presence on the internet with a website;
- Online stores (e-commerce), e.g., Amazon and Zalando;
- Online services, e.g., AirBnB, eBay, Booking.com, and dating sites;
- Internet-related information and communication technology (ICT) such as web design, hosting, and internet marketing.

Estimating the size of the internet economy is not easy. Only the last group, internet-related ICT, is to some extent distinguished as a separate industry in statistical surveys. Online stores are, according to European rules, grouped in a single industry, so that there is hardly any information about the type of their activities. Online services are subsumed in the industries corresponding to their output, so that, again, there is hardly any information about the type of their activities. And about the internet as means of business communication there is no regular statistical information at all. By combining big data, administrative data and survey data, Statistics Netherlands (Oostrom, 2016) has succeeded in estimating the size of the internet economy for each of these four classes. The data consisted of:

- *Big data*:
  List of all Dutch websites, including business name, company-registration number, size of site traffic, and other data. The list contains 2½ million websites and is maintained by Dataprovider (https://www.dataprovider.com). When a website is owned by a company, it is legally obliged to publish its company-registration number.
- *Administrative data*:
  General Business Register (GBR): a comprehensive list of all enterprises in the Netherlands and their ownership relations, based on administrative data from the company register and the tax registers.[2] There are 1½ million enterprises in the GBR. Addresses and telephone numbers are also included in the GBR; for many enterprises, a hostname of the website is included.
- *Short-term statistics (STS)*:
  turnover of enterprises based on the VAT (value added tax) register, giving a nearly complete census of enterprises.
- *Survey data*:
  Structural Business Statistics (SBS): annual survey of enterprises on employment and financial data; all enterprises with more than 50 employed are included, smaller enterprises are sampled.
- Several other surveys such as on ICT and wages.

Several keys were used in linking all these databases: the company-registration number, address and telephone number, and hostname.

The statistical results were remarkable. About two thirds of all enterprises have no website; most of these are self-employed persons. Online stores, online services, and internet-related ICT, are in size about 5 per cent of the economy, comparable to construction and transportation. So the internet economy is important, but its size is not very large, much smaller than for example health services and education

[2]An enterprise in official statistics is more or less the same as a business unit, and a legal unit is an entity officially registered as an economically active unit, such as a corporation and a self-employed person, whereas an enterprise group (a set of enterprises controlled by the same owner) is more or less the same as a company. See Eurostat (2016) for more details and references to implementation rules.

services. On the other hand, the use of ICT and internet is probably very extensive, since most enterprises use computers and the internet.

6.4.2 Big data and the quality of GNP estimates

Big data may also contribute to the improving the quality of gross national product (GNP) estimates. This is important, because there is a need to improve the quality of the first estimate of quarterly GNP, 45 days after the quarter (Zeelenberg, 2017). Also, in many countries users need an even more timely first estimate, preferably 30 days after the end of the quarter.
There are four possible big data sources, here:

- *Tax databases*: These are mainly for taxes on wages and on sales. These are not yet rapid enough. They lag about two months: one month for reporting taxes and one month for filling the databases.
- *Company accounts*: It will become possible within the next few years to have direct access to company accounts. Direct access in the sense that reporting modules for taxation and statistics will have been built in the accounting software. But this is not yet possible, and even when it will have been implemented, then only for annual accounts.
- *Banking transactions of enterprises and households*: This is clearly the most promising. There is only one clearing-house for banks in the Netherlands, and it is very rapid. Also, banks have to report daily, weekly and monthly to the central bank. But there are of course very strong privacy concerns here. So it will be very, very difficult, and will take a long time before even the first steps will be taken and even longer before it will be implemented.
- *Model-based estimates based on big data*: not a big data source as such, of course, but a way to use big data. At the moment, this appears to be the best option.

An example of what model-based estimates from big data may have to offer, is given by Van Ruth (2015) who analyzed the relation between traffic intensity and economic activity in an important region of the Netherlands, around the city of Eindhoven. Traffic intensity is measured from traffic sensors in the road surfaces, and economic activity is measured by expected output, taken from the monthly Manufacturing Sentiment Survey. The traffic intensity indicator tracks that of expected output amazingly well. Peaks and troughs coincide, meaning that the traffic intensity index should be able to signal important turning points in economic activity. Statistically, the series appear to be coincident, and possibly seasonal adjustment and a trend-cycle decomposition may remove some noise and further improve the model. Now, traffic intensity data becomes available with a much shorter time lag than traditional survey data. From the model that relates output to traffic intensity, we may then make a preliminary estimate of output, which in turn may be used to improve the first estimate of GNP. So, this and similar models, might be useful in making better first and preliminary estimates of GNP.

6.4.3 Google Trends for Nowcasting

Choi & Varian (2011) 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 even 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 (Lazer et al, 2014). 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 (EU, 2005), is that NSIs should warn the public and policy makers when statistical results are being used inappropriately or are being misrepresented. As emphasized by Reimsbach-Kounatze (2015) and Fan, Han & Liu (2014), 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 analyzing big data.

7. Analytical use of big data

In recent years, new techniques, usually grouped under the heading machine-learning techniques, have been developed that may be used to analyse big data. These techniques have been developed for high-dimensional data (i.e., data with a large number of variables per record), whereas the big data we have discussed above often have a low, sometimes even very low, dimension. However, some experiments have suggested that within official statistics, they may be useful also for low-dimensional data, for example to correct for selectivity.
Also, for statistical analysis, i.e., for answering what-if questions, it is not always necessary to have representative data, and so big data might be useful for analysis. But insight into the data quality and the data-generating process is always necessary, and, as we have seen above, this is a problem for many big data from intermediaries.
It is crucial to remember that any use of models must be made explicit. It should be documented and made transparent to 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 making purely behavioral 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.

8. Discussion

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. An important caveat is that often a big data source does not cover the entire target population that is interesting from a statistics user's perspective. This may introduce selectivity in statistical estimates which limits the usefulness of big data from a statistical perspective. A number of possible solutions and workarounds have been discussed in this chapter but more work should be done in this important area.
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.
It is crucial that NSIs continue to actively explore opportunities of big data. Many more sources will emerge, and may become available for production of statistics. In the near future, biological data, e.g., on genomes, and medical data, on health and care of individuals, will become available for scientific research and for linking with social data on income, crime, jobs, etc.[3] The internet of things, consisting of all kinds of large and small devices, is expected to generate a tremendous amount of data from which e.g., information on personal behavior can be derived: movement patterns, health aspects, energy consumption, and much more. Sensors in the public space will provide information on the environment like air quality or noise pollution, contributing to the concept of a smart city. Smart manufacturing and smart agriculture refer to industries that take advantage of intensive generation and analysis of large amounts of data. And in addition to physical space, virtual or cyber space becomes an increasingly important study object in itself, giving rise to new phenomena like cybercrime or the internet economy. The exploration of opportunities will be accompanied by non-trivial challenges as argued in this chapter, but NSIs are in an excellent position to use their traditional experience and high quality standards in innovative ways.

[3] The Royal Netherlands Academy of Arts and Sciences has placed a proposal for such a database by universities and Statistics on its *Academy Agenda for Large-scale Research Facilities* which need to be in place by 2025. This Agenda lists research facilities that could produce new scientific breakthroughs (KNAW, 2016).

9. References

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

Braaksma, B., & K. Zeelenberg, 2015, "Re-make/Re-model": Should big data change the modelling paradigm in official statistics? Statistical Journal of the IAOS 31, 193–202. doi: 10.3233/SJI-150892

Buelens, B., de Wolf, P.-P., & Zeelenberg, K., 2017, Model-based estimation at Statistics Netherlands. Discussion Paper, Statistics Netherlands, The Hague.

Choi, H., & Varian, H. R., 2011, Predicting the present with Google trends. http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf

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

Daas, P. J. H., Puts, M. J., Buelens, B., & van den Hurk, P. A. M., 2013, 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

De Jonge, E., van Pelt, M, & Roos, M., 2012, 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

De Meersman, F., Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A., Demunter, C., Reis, F., & Reuter, H. I., 2016, Assessing the Quality of Mobile Phone Data as a Source of Statistics. Paper presented at the European Conference on Quality in Official Statistics, Madrid, 2 June 2016. http://www.ine.es/q2016/docs/q2016Final00163.pdf

De Wolf, P.-P., & K. Zeelenberg, 2015, Challenges for statistical disclosure control in a world with big data and open data. Invited paper for the 60th World Statistics Congress. http://www.isi2015.org

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

EU (European Union), Regulation on European statistics, 2009, Official Journal of the European Union, L 87 (31 March 2009), 164–173, http://data.europa.eu/eli/reg/2009/223/2015-06-08

Eurostat, 2016, Statistics Explained - Glossary: Enterprise. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Enterprise

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

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

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

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

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

KNAW(Koninklijke Nederlandse Akademie van Wetenschappen: Royal Netherlands Academy of Arts and Sciences), 2016, Thirteen selected facilities and three honourable mentions. https://www.knaw.nl/en/advisory-work/copy_of_knaw-agenda-grootschalige-onderzoeksfaciliteiten-13-geselecteerde-faciliteiten?set_language=en

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

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

Oostrom, L., Walker, A. N., Staats, B., Slootbeek-Van Laar, M., Ortega Azurduy, S., & Rooijakkers, B., 2016, Measuring the internet economy in The Netherlands: a big data analysis. Discussion Paper 2016-14, Statistics Netherlands, Heerlen. https://www.cbs.nl/nl-nl/achtergrond/2016/41/measuring-the-internet-economy-in-the-netherlands

Reimsbach-Kounatze, C., 2015, The proliferation of “big data” and implications for official statistics and statistical agencies: a preliminary analysis. OECD Digital Economy Papers 245, OECD, Paris. doi:10.1787/5js7t9wqzvg8-en

UN (Statistical Commission of the United Nations), 1991, Fundamental Principles of Official Statistics, revised edition, 2014, http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

Struijs, P., Braaksma, B., & Daas, P. J. H., 2014, Official statistics and big data, 2014, big data & Society, April–June 2014, pp 1–6, doi: 10.1177/2053951714538417

Telecompaper, 2015, Majority of the elderly in the Netherlands has a smartphone. Press release at https://www.telecompaper.com/pressrelease/majority-of-the-elderly-in-the-netherlands-has-a-smartphone--1088067

Tennekes, M. and Offermans, M.P.W., 2014, Daytime population estimations based on mobile phone metadata. Paper prepared for the Joint Statistical Meetings, Boston.

http://www.amstat.org/meetings/jsm/2014/onlineprogram/AbstractDetails.cfm?abstractid=311959

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

Vaccari, C., 2016, Big Data in Official Statistics. Saarbrücken: Lambert.

Van Ruth, F. J., 2014, Traffic intensity as indicator of regional economic activity, Discussion paper 2014-21, Statistics Netherlands, 2014.

Varian, H. R., 2014, Big data: new tricks for econometrics. Journal of Economic Perspectives 28(2), 3-28, doi: http://dx.doi.org/10.1257/jep.28.2.3

Zeelenberg, K., 2016, Challenges to Methodological Research in Official Statistics. Presentation at the 2016 International Methodology Symposium, 22 March 2016, Gatineau. http://www.statcan.gc.ca/eng/conferences/symposium2016/program

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/'18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/'14–2017/'18	Crop year, financial year, etc., 2015/'16 to 2017/'18 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information