Quality Approaches to Big Data in Official Statistics

Peter Struijs and Piet Daas, Statistics Netherlands[1]

It is widely recognised that important methodological and quality issues are associated with Big Data. Especially selectivity issues become prominent when trying to apply established statistical methods. Using sampling theory as the framework for the production of estimates based on Big Data may not be effective, especially if the data cannot be linked to a known population of units. The question arises to what extent an approach originally developed for survey based statistics can be applied to Big Data. The paper discusses possible quality approaches to the production of official statistics when dealing with Big Data, for instance the compilation of statistics not aimed at predefined populations. These approaches could result in rapid information with high relevance from a user's perspective. However, the introduction of such approaches requires an assessment of the role that National Statistical Institutes aspire to play in the era of Big Data.

1. Introduction

According to a Task Team of the UNECE High-Level Group for the Modernisation of Statistical Production and Services, Big Data (BD) can be defined as "data sources that can be – generally – described as high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making" [1]. This paper explores what the use of such data sources imply in respect of statistical quality and methods.

Quality and methodology are closely related. The quality of statistics depends on the methods applied, and methods are usually chosen to fulfil quality objectives. A large part of the body of established methods is connected to sampling theory, the core of which refers to a target population of units and variables, to which sampling, data collection, data processing and estimation are tuned and optimised, considering cost and quality aspects. However, as we will

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the position of Statistics Netherlands. The e-mail addresses of the authors are p.struijs@cbs.nl and pjh.daas@cbs.nl, respectively.

see, it may be far from obvious whether or how BD usage can be related to this. In fact, there are a number of methodological issues related to the use of BD, as recognised in the description of the ongoing BD project of the UNECE [2] and by the DGINS in their so-called Scheveningen Memorandum [3]. This leads to three questions discussed in this paper:

- What are the limitations of the established quality framework and methodology for official statistics based on statistical surveys and administrative data sources, when BD is used, whether as a primary data source or in a more supportive role?
- Considering these limitations, what are the alternatives for dealing with the challenges?
- Considering the limitations and the alternatives for dealing with them, what choices should be made, taking the emerging future environment of NSIs into account?

Since these questions are too "big" to be answered in depth in a short paper, we aim at presenting a line of thought for consideration.

2. Quality approaches and challenges

2.1 Established methods

With only very few exceptions, the statistical programmes of NSIs are based on inputs from statistical surveys and administrative data sources. For such statistics there exists an elaborate body of validated statistical methods. Many of these methods are survey oriented, but in fact, most survey based statistics make use of population frames that are taken or derived from administrative data sources. Methods for surveys without such frames do exist, for instance area sampling methods, but nowadays even censuses (of persons and households as well as business and institutions) tend to make use of administrative data. And administrative data sources are, of course, themselves also used as the main data source for statistical outputs. Statistical surveys are increasingly used for supplementing and enhancing administrative source information rather than the other way round. This is the consequence of the widely pursued objectives of response burden reduction and efficiency.

In addition to so-called stovepipe statistics, which are compiled in processes that are run more or less independently from each other, there also exist integrative statistics, based on a multitude of

sources. The prime example of such statistics is National Accounts (NA). Statistical methods for NA focus on the way different sources for various domains and variables of interest can be combined. Since these sources may be based on different concepts and populations, frames and models have been developed for integration of sources. These frames and models include, for instance, macroeconomic equations. Interestingly, NA outputs generally do not include estimations of business populations. This may reflect the fact that the production of NA involves quite a few expert assumptions as well as modelling, rather than population based estimation.

This characterisation of statistical methods is, of course, incomplete. There are a number of methods aimed at specific types of statistics, for instance occupancy models for estimating the evolvement of wild animal populations, or time series modelling. Especially methods outside traditional sampling theory may be interesting when dealing with BD.

2.2 Established quality approaches

What role do quality considerations play in official statistics? Two levels can be distinguished. First, the quality of statistics is the result of the methods applied – and their parameterisation. Methods are judged and the parameterisation chosen on the basis of their effect on quality, mostly with a focus on accuracy. Second, quality frameworks have been developed that apply to entire statistical programmes, which provide criteria for judging these programmes. Examples are the European Statistics Code of Practice (CoP) [4] and the Quality Assessment Framework (QAF) [5], for which specifications exist for certain domains, including NA. The CoP defines five aspects of statistical output quality:

- relevance
- accuracy and reliability
- timeliness and punctuality
- coherence and comparability
- accessibility and clarity

In this paper we focus on the first four aspects, and ignore the aspect of comparability over time, since a related paper of this conference discusses the latter [6].

2.3 Big Data challenges: examples

Let us now turn to BD and look at three examples from methodological research at Statistics Netherlands. What methodological and quality issues were encountered that required innovative solutions?

The first example concerns the use of information from about 20.000 sensors on Dutch roads, from which the number of passing vehicles in various length classes is available on a minute-by-minute basis [7]. This source has the potential to be used in statistics on weekly traffic indices including specifications for heavy traffic, with a detailed regional breakdown. Issues encountered include:

- The areal distribution of road sensors is uneven, there are lapses, and at the lowest time period not all data are available for all sensors.
- The relationship between the population of vehicles and the road sensor data is not known at the micro-level. Nor can individual vehicles be traced through time.
- The metadata of the road sensors has a poor quality.

The second example is the use of public social media messages, such as Twitter or Facebook, which has the potential to be used for sentiment indices on a weekly basis, including a consumer confidence index [8]. Issues encountered include:

- The population behind the messages is not known, and neither is its relationship with the population at large.
- It is possible to devise a system for attributing a sentiment to text messages, but it is not obvious how to anchor or interpret the sentiment thus measured.

The third example is about the use of mobile phone location data [9]. This has the potential to be used for statistics on where people are at any specific time, so-called daytime population statistics, or for mobility and transport statistics. Tourism statistics may also use this data source. Issues encountered include:

- The data available depends on the measurement grid of the mobile phone provider and its robustness.

- Even if data is available on the owner of mobile devices, the phones may not be used, be switched off or used by other persons than the owner.

2.4 Big Data challenges: towards an overview of issues

Three examples are not enough for identifying all BD challenges regarding methodology and quality. However, we must keep in mind that, although research on the usability of BD for official statistics is now taking off, there are still hardly any official statistics that are actually based on BD sources. So for the time being we have to do with the challenges that are popping up in this research.

Ideally, one would look at all types of BD that occur and consider what methodological and quality impediments exist for its use for official statistics. A typology of BD sources has been considered in the BD project of the UNECE mentioned earlier [10]. Statistics Netherlands has developed a BD roadmap, which it had assessed by IBM [11]. The assessment was based on an IBM typology of BD sources, which was linked to statistical domains. These typologies and the assessment itself make clear that there are far more BD sources – and types of BD sources – than actually considered for use in official statistics. So it is clearly too early to list all BD challenges.

Nevertheless, it may be useful to mention the main types of methodological and quality issues that have been encountered so far. These include:

1. Information about the population behind the records used may be missing. This may occur at the micro-level (making micro-linkage impossible) and at the macro-level (lack of information on selectivity).
2. The measurement grid may show an unbalanced or unstructured physical distribution, may contain gaps or may suffer from other types of under- or overcoverage. There may also be coverage issues along the time dimension.
3. The meaning or relevance of the data may be difficult to gauge. What information is actually conveyed by a text message, a sentiment, an entry to a search machine or a photo?

In addition, there are other types of methodological and quality issues, not discussed in this paper, that result from technical and cost limitations that have to do with processing the enormous volume and velocity of BD [12]. But, although traditional statistical methods may fall

short, this does not mean that the higher-level quality framework cannot be applied anymore. In particular relevance, reliability, timeliness and coherence are quality requirements that may still be applied when dealing with BD.

3. Ways of dealing with the challenges

What courses of action can be considered when confronted with issues for which there are no established methods available? It should be kept in mind that, in theory, BD may be used as a single source for statistical outcomes, or BD sources may be combined with statistical surveys or administrative data sources. The challenges depend on the envisaged type of use, of course, but if they cannot be solved by the use of traditional methods, the following alternatives can be considered (see also [13, 14]):

1. Missing information about populations:
    - In some cases there are ways to derive at least some information about the population behind BD. For instance, for social media messages it may be possible to estimate background variables based on the correlation between the wording of the message and the age, gender or social group of the messenger. Information on background variables may then allow the application of established methods.
    - In other cases BD may be related at the meso- or macro-level to other information, allowing modelling approaches. For example, even if the population behind mobile phones is not known, its relationship with administrative population registers can be studied at an aggregated level. Mobile phone movements can be related to existing traffic statistics, and so on.
    - The approaches mentioned below (3. Meaning and relevance issues) may also be considered.
2. Measurement grid and coverage issues:
    - For many issues concerning the measurement grid and coverage, established methods can be adapted to the need, but there are situations that call for modelling approaches that do not belong to the standard toolkit of statisticians, such as probabilistic modelling and the use of occupancy models. The knowledge and experience of NA may also inspire solutions to coverage issues.

3. Meaning and relevance issues:
    - If there are difficulties in understanding the meaning of certain BD, it might be possible to study its relationship with data from other sources with which a stable correlation seems plausible. Indicators based on the BD set can then be calibrated or fitted to the other data set. For instance, a sentiment index based on social media data by applying text mining can be fitted to an already existing survey based consumer confidence index [8]. The BD source can then be used for producing rapid – or, to remain on the safe side, provisional – consumer confidence figures.
    - Even without fitting, correlations with other, known, phenomena can be produced. The stability of correlations can be substantiated if BD allows successful forecasting. Even stable correlations do not imply causal relationships, of course, and by producing information on correlations the issue of meaning is not really solved, but it would allow users of the information to judge for themselves.
    - The most radical approach would be to produce BD based information as "stand-alone", leaving the interpretation completely to the users. At first sight this may look like a silly proposition for an NSI, but there may be demand for new types of information that do not have an obvious interpretation. For instance, there is unmistakably demand for Twitter mood data [14]. If such information is presented as a "beta-version", it might be given a chance and elicit valuable feedback from users. Some internet "giants" heavily promote these kind of approaches.

One could also think of approaches that change the setting of the problem. One could tackle the selectivity issue associated with unknown populations by launching a survey to collect characteristics on such populations, for example the users of Facebook. After such a survey estimation methods can be readily used [15]. This is reminiscent of using surveys to measure the quality of administrative data sources. Another approach, simply lowering quality requirements, may seem undesirable, but if BD is used for provisional figures, this may be acceptable.

4. Making the right choices in a changing environment

When considering the possible use of BD sources and seeing the issues, what approach should an NSI take? There are many reasons why one might be very cautious or even wary of embracing BD. One of the main assets of NSIs is public trust, loosing this must be risked under no circumstances. Providing information on spurious correlations or on phenomena described by BD that are not well understood is not the task of NSIs. Official statisticians have to be held to the highest professional standards. Moreover, BD is a hype not to be followed blindly, it may be better to wait and see. The information yielded by BD that catches headlines is anyway not all that convincing, as we know from the discussion about Google Flu Trends [16]. And we have not even considered non-methodological issues such as privacy and the public image of NSIs, or IT challenges. Isn't the business case for using BD for official statistics clearly a negative one?

No, it is not.

Consider the following. The environment in which official statistics are being produced is changing. Traditionally, the production of official statistics was a monopoly. It still is, but where in the past virtually the only statistics on social phenomena available to the public were official statistics, now non-official statistics are rapidly becoming widely available [17]. Their quality and objectivity may be disputed, but they are there, they are produced and disseminated much more rapidly than official statistics, and they are used. There is a real risk that this will erode the position of NSIs, in particular their funding. It is true that official statistics still fulfil a vital role in society. NSIs may, correctly, point to the fact that in a society that is overwhelmed by information provided by a host of providers, an impartial anchor – official statistics – is crucial. But there is no guarantee whatsoever for the survival of NSIs in the long run.

There is reason for NSIs to go back to first principles and see what role they aspire to play in the era of BD. The availability to society of impartial information of good quality according to need has to be ensured, that is fundamental. But there is no intrinsic necessity for this information to be produced by NSIs. Others may produce the information, if this is validated by NSIs. There will probably remain a core set of statistics that others than the NSIs will not produce, and NSIs need to maintain their knowledge position in order to carry out their validation role. In this way the trust in NSIs is used as an asset and will be reinforced.

What does this mean for the use of BD by NSIs? One could think of the following:

- NSIs must have or must obtain knowledge and experience of how BD can be used and cannot be used. Knowledge is also needed about how BD is used outside NSIs. The principle of "quantity over quality", adopted by BD users such as Google, should not be discarded out of hand.
- Even if BD is not used for new statistical outputs, BD may be used in the existing programme of official statistics if this leads to efficiency or response burden reduction, provided the challenges can be overcome.
- The use of BD for compiling early indicators for important statistics, such as price statistics or business cycle statistics, is a serious option. The use of BD for nowcasting can also be considered.
- The traditional way of designing statistical processes is to define the desired outcome, choose suitable data sources and optimise the process. Experiments with BD may be carried out where this approach is turned around: take an interesting BD source, start compiling information that may be relevant and then try to relate that information to information already available, if only by establishing correlations.
- It is necessary to create an institutional environment in which experiments with BD can be carried out. This has to do with IT, with HRM, with strategic backing of BD initiatives and with openness to non-conventional solutions, such as forecasting. A mind-set is required in which data sources are not just looked at in terms of their "representativity".

Taking BD at face value and studying its relationship with other phenomena may not be as outlandish as it may seem. When administrative data sources are used by NSIs, they are sometimes also taken at face value, instead of being used as a source for measuring pre-defined statistical concepts. For instance, users may be interested in the occurrence of crime, but there are NSIs that produce statistics of reported crimes instead, based on police registers. In fact, taking data from any source at face value is relatively low-cost, be it administrative data or BD.

The link between correlation, causation and forecasting was mentioned earlier. This is an area where economists and econometrists have ample experience [18]. If this becomes an important work area for NSIs, there may be reason to reconsider the institutional boundaries of NSIs. INSEE, the French NSI, for instance, also has a mission regarding economic research. One may

wonder whether the traditional distinction between official statistical and official economic and other forecasting institutes that exists in many countries still needs to be maintained.

Following the ideas presented above will result in another approach to quality. Statistics will still be of high quality, in the sense of being fit for purpose, and produced according to the highest professional standards. The core elements of quality, that is, relevance, reliability, timeliness and coherence, will remain of the utmost importance in the BD era. But their contents will evolve, together with the role of NSIs. The same will be true for the professional standards. In fact, the quality approach to BD in official statistics as described will amount to a paradigm shift.

References

[1] UNECE (2013), What does Big Data mean for Official Statistics?, paper produced by a Task Team on request of the High-Level Group for the Modernisation of Statistical Production and Services, available here.

[2] UNECE (2013), The role of Big Data in the modernisation of statistical production, project plan for 2014 as approved by the High-Level Group for the Modernisation of Statistical Production and Services, available here.

[3] DGINS (2013), Scheveningen Memorandum on Big Data and Official Statistics, available here.

[4] ESSC (2011), European Statistics Code of Practice, available here.

[5] ESSC (2012), Quality Assessment Framework, version 1.1, available here.

[6] Booleman, M. et al (2014), Statistics and Big Data: Quality with uncontrolled inputs, paper prepared for the Q2014 conference.

[7] Daas, P.J.H., Puts, M.J., Buelens, B., Van den Hurk, P.A.M. (2013), Big Data and Official Statistics, paper for the 2013 NTTS conference, Brussels, Belgium, available here.

[8] Daas, P.J.H., Puts, M.J.H. (2014), Social Media Sentiment and Consumer Confidence. Paper for the Workshop on using Big Data for Forecasting and Statistics, Frankfurt, Germany, available here.

[9] De Jonge, E., Van Pelt, M., Roos, M. (2012), Time patterns, geospatial clustering and mobility statistics based on mobile phone network data, discussion paper 201214, Statistics Netherlands, available here.

[10] UNECE (2014), How big is Big Data?, the role of Big Data in Official Statistics, version 0.1, draft for review, prepared by the UNECE Task Force on Big Data, available here.

[11] IBM (2014), Big Data roadmap assessment, carried out on behalf of Statistics Netherlands.

[12] Daas, P.J.H., Puts, M.J. (2014), Big Data as a Source of Statistical Information. The Survey Statistician 69, pp. 22-31, available here.

[13] Struijs, P., Daas, P.J.H. (2013), Big Data, big impact?, paper presented at the Seminar on Statistical Data Collection of the Conference of European Statisticians, Geneva, Switzerland, available here.

[14] Bollen, J., Mao, H., Zeng, X-J. (2011), Twitter mood predicts the stock market, Journal of Computational Science, 2(1), March 2011, available here.

[15] Buelens, B., Daas, P., Burger, J., Puts, M., Van den Brakel, J. (2014), Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, available here.
[16] Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014), The parable of Google Flu: traps in Big Data analysis, Science, 14 March 2014, available here.
[17] Struijs, P., Braaksma, B, Daas, P.J.H (2014), Official Statistics and Big Data, to be published in Big Data and Society, announced here.
[18] Varian, H.R. (2014), Big Data: new tricks for econometrics, Journal of Economic Perspectives, volume 28-2, available here.