**A statistical approach to Big Data**

Gustav Haraldsen, Statistics Norway, Gustav.Haraldsen@ssb.no

Arild Langseth, Statistics Norway, Arild.Langseth@ssb.no

**Abstract**

*Statistics Norway is presently working on a strategy on how to utilize new data sources. We argue for a strategy which combines what we name a challenge-based approach and an opportunity-oriented approach. The challenge-based approach takes quality issues recognized by the statistical producers as a starting point, and we report from three focus groups that we have conducted addressing this. The opportunity-oriented approach takes the characteristics of Big Data as a starting point and explores what kind of statistics which can be produced from them. We discuss three main ways to produce statistics from Big Data sources.*

***Keywords:*** *Statistics, Big Data, new data sources*

**1. Introduction**

A statistical approach to Big Data should reflect the unique characteristics and purpose of statistics. We would like to emphasize the need for stable data sources, containing valid and reliable data about main social and economic processes. Generally, data validity, which is how well the data collected represent the population it was meant to cover and measure what the questionnaire were meant to measure, is the most challenging part of data collection. Moreover, statistics is about comparing aggregates in time and between classified entities (like social groups or geographical areas). Consequently, breaks in time series or definitions cause crises in statistics and call for explanations and adjustments. Some breaks are inevitable, caused by social changes. But others may be caused by changes in data sources or data collection instruments.

**2. Big Data Characteristics**

The term "Big Data" mainly refers to the big volume of data, often said to be so big that traditional software cannot handle it. In addition, its velocity and variety are included in the original definition. Later veracity and value have also been added (Mayer-Schönberger 2013). The problem with all these characteristics is that what falls inside or outside the definition will change as software and other IT technology change (Few 2018). Some analysis even suggests that the original 3Vs are not necessarily defining characteristics of Big Data (Kitchin 2016). The social statistician Robert Groves has suggested the alternative term “organic data”, because we are talking about data which reflects how the social organism is working (Groves 2013). We tend to favour this approach. For the purpose of this paper we define Big Data as digital traces of human activities, actions and attitudes (César Hidalgo referred to in (Few 2018)). Instead of asking people what they do and think, these new data sources open the possibility to produce statistics based on the digital “fingerprints” left behind by people’s actions and attitudes; be it process data or data about products, services or other consequences of human actions (Salganik 2018). Using this definition as a starting point, we can focus on how social and economic behaviour are represented in digital sources. Looking at how technology also affects our behaviour and attitudes will be a future step.

In addition to having some common characteristics, different kinds of Big Data sources also have their own characteristics. We distinguish between three main kinds of data (UNECE 2013):

1. Process-mediated data, which includes transactions and results of transactions made by businesses and governments. These are the most direct measurements of human actions, but also often some of the most protected kind of Big Data.

2. Machine-generated data, often called The Internet of Things, which are traces of behaviour or products of behaviour registered by sensors, cameras or software built into machines. Note that machines may indicate, but not always tell the whole story about what people are doing. A good example is the TV-meters that were introduced back in the 1980’s. They recorded which television channel people had on, but not if they were watching.

3. Human-sourced information or opinions communicated by broadcasting or in social media. The main focus has been on social media. But because this information and these opinions are initiated for a certain purpose, they may not give a representative picture of what people think and believe in (Couper 2013).

The potential and challenges of incorporating Big Data into statistics are both affected by the common characteristics of Big Data, as well as the particularities of different kinds of Big Data.

**3. Challenge-based and Opportunity-oriented approach**

New data sources can address current challenges in present data collections or statistics, or open new, unprecedented opportunities for statistics. An approach for utilizing these digital sources will therefore be on a scale that goes from what we can call a strictly challenge-based approach to a completely open, opportunity-oriented approach. While the challenge-based approach springs from statistical and methodological needs for improvements, an opportunity-oriented approach uses the qualities of the different digital data sources as a starting point and explores what kind of statistics which can be produced from them.

We believe that the challenge-based and the opportunity-oriented approaches can be combined. Furthermore, we think the challenge-based approach should be organized as concrete projects initiated by statistical or methodological units, while the opportunity-oriented approach should be explored by an innovative group of methodologists, researchers, computer scientists and domain experts, recruited according to an internship scheme. To maintain continuity, coordinate the work done, look for new opportunities and build competence, we envisage the need for an overarching Centre of Data Expertise with cutting edge knowledge and ideas on how to extract valuable information from data. Figure 1 illustrates this combined approach.

**Figure 1. Challenge-based and Opportunity-oriented approach**



**4. Improvement to current statistics**

To identify statistical challenges that may be addressed by new, digital data sources, we conducted three focus groups with heads of subject matter divisions. The focus group agenda was quality issues and statistical needs that are not sufficiently covered by the present data sources.

The discussion revealed that incomplete data sources was considered to be a major challenge. It is a well-known problem that nonresponse in voluntary surveys often causes certain groups to be poorly represented. What is not so often discussed is that many of the administrative registers also are incomplete. Examples of services not covered by the registers are private health services, education offered on the Internet and informal services like Airbnb and Über. What is recorded in administrative registers is also to some extent affected by the political agenda. Criminal records are one example. Both the nonresponse problem in surveys and the problem with incomplete registers are increasing.

By replacing survey questions with new digital data, we reduce response burdens. It was pointed out that this is particularly important in business surveys, where time is money, and in personal surveys with detailed questions about consumption, daily activities or travelling. A lower response burden in voluntary surveys may also contribute to a higher response rate.

The focus groups also brought up the need for more direct measurements of economic transactions. One example was the need for actual purchase prices of goods instead of commodity prices from one source, which is adjusted for shopping patterns gathered among different kind of customers from another source.

It was pointed out that more complete datasets may offer interesting opportunities for more detailed statistics within health statistics, immigration statistics, innovation statistics and others. Additionally, more frequent statistics are wanted in some areas, in particular in economic statistics. One should remember, however, that statistics first and foremost should identify trends and not short-term, short-lived events. An excessive emphasis on day-to-day statistics may leave the impression that society is changing faster than what is the case.

A kind of real time statistics that could be very useful, however, is statistics about our own data collections. Our surveys use digital technology and therefore leave large amounts of digital traces that can be used to analyse and improve the recruitment and response processes, preferably during the data collections (what is commonly called a responsive approach). In this way we can use Big Data tools to study our own data collection and improve our own work processes.

Figure 2 gives an illustration of how a quality indicator based on the number of activated error messages (Haraldsen 2005) can be followed during the data collection and summarized in a box-plot when the data collection is completed.

**Figure 2. Real time quality measurement**



**5. Data design methods**

Traditional data collections are designed to collect data relevant to a certain statistics from a population or sample of identifiable units. The delivery is a data matrix with unit identifications at the beginning of each line and the different data variables along the following columns. This matrix will be the basis for statistical analysis.

In designed data collections we formulate a set of questions which are intended to collect certain kinds of information. When we capture data from secondary sources it is the other way around. We approach a set of given data with some questions. This procedure calls for a content analytic approach, similar to the approach taken by media scientists when studying patterns and trends in media content. Content analysis is commonly divided into three steps; Extract data, clean, rearrange and code them into predetermined categories and eventually load the results into the data matrix (Struhl 2015).

To make practical use of Big Data is even more complicated. Partly because of the sheer volume of data, and more significantly the volume of noise relative to meaningful data. Cleaning, rearranging and coding is also more challenging. Graphical representation, machine learning and Bayesian modelling are used to handle these challenges. A more fundamental issue, however, is that the captured data seldom fit into a traditional data matrix. Persons, enterprises or other units may not be identifiable, and there are few clues on how to interpret the data. In figure 3 we have suggested three main ways to meet this challenge. They are named Generating matrices, Direct classification and Combined modes.

**Figure 3. Alternative ways of producing statistics**

****

*5.1 Generating matrices*

The first and most obvious option is to look for ways to produce the same kind of data matrix that we would normally use for statistics. One example is Statistics Norway’s effort to link receipts from grocery stores with account details from banks.

In the Household Consumer Expenditure Surveys participants are asked to record everything bought over a fortnight period. To specify grocery items is the most burdensome task and the response rate in these surveys are unacceptably low. If we can link grocery stores receipts to bank accounts, this task can be avoided. And we can. We do this by linking the timestamp on the receipt with the timestamp on the bank transaction in combination with the name of the store. The owner of the account used for the transaction is identified by his or her personal identification number (PIN). From this a traditional matrix with identifiable customers and goods bought can be constructed.

In the grocery example we need the PIN because we want to study grocery expenditure in different parts of the population. Likewise, in economic statistics we use the legal organisation number of enterprises and establishments to classify different kinds of businesses. These are the two most common identifiers. But there are alternatives, for instance the use geographical locations, and some of these are both easier to find in Big Data and interesting to use in statistics.

*5.2 Direct classifications*

Identification numbers are only a classification tool in statistics. As soon as the units are classified, data can be anonymized. A different approach when identifiers are missing or unavailable is therefore to look for alternative classification methods. Identifiers may exist but be unavailable because we do not have access to them. In these cases, we may be able to have data classified without us seeing how it is being done. A simple example is when we collect statistics prepared by other organizations. A more advanced version is when data from different sources are linked together and classified by computer driven algorithms that do not disclose identifiable units.

One computer driven classification method that are used by Google, Facebook and other commercial Internet companies is to derive classification characteristics from digitalized patterns of behaviour. One simple example on how this technique can be used in statistics is that daily travel routes recorded by position data can reveal:

• if you have small children (because you regularly go to kindergartens)

• if you are a full-time worker (because you normally go to and leave the same workplace each weekday)

• and what kind of branch you are working in (because we know what is produced at different sites).

*5.3 Combined methods*

The final alternatives listed in figure 3 are different kinds of combinations between survey data and Big Data. The traditional multi-mode methods in surveys are either to offer different ways to report at the same step of the data collection (mixed-mode) or different ways to report at different steps of the data collection (multi-mode). We suggest a third alternative, named combined modes. Combined modes are when different modes are used for different purposes. Surveys will often give better estimate of levels, while Big Data can be a source of trend or forecast estimates. As an example, survey data may be used to estimate the number of tourists visiting Norway or Norwegian shopping across the border at a certain time. The result is then compared with foreign mobile activity, credit card payments or other activity indicators gathered from digital sources. Later, changes in the activity patterns are used to estimate how the number of tourists or customers change (e.g. (Desamparados 2019)).

The last option listed in figure 3 is when survey results are used to evaluate and adjust results collected from alternative data sources. The term “audit” is commonly used when an independent evaluator is, for example, examining the quality of services offered or product produced by others. When surveys are used to audit data from digital sources, focus will be on how representative, valid and reliable the results are, and possibly how the quality can be improved. In other words, the survey is used as a gold standard and source for adjustments. Considering that Big Data often is presented as a replacement that will take over for surveys, this is truly a paradox.

**6. Conclusions**

The main characteristics and quality aspects of statistics need to form the foundation for any statistical attempt to utilize new data sources. We propose a combination of a challenge-based and an opportunity-oriented approach to incorporate Big Data sources into statistics. The former as concrete initiatives to improve current statistics. The latter as innovative explorations to find new methods to classify and combine data, as well as ways to cope with specific Big Data shortcomings or quality issues.

**7. References**

Mayer-Schönberger V. and Cukier K. (2013), Big Data. A Revolution that will transform how we live, work and think, John Murray, London, UK.

Kitchin R. and McArdle G. (2016), What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets, Big Data & Society, Sage Publishing, California.

Groves R. (2013), Official statistics and big data, NTTS Conference, Brussels, Belgium.

Few S. (2018), Big Data, Big Dupe. A little book about a big bunch of nonsense, Analytics Press, El Dorado Hills, California.

Salganik M. (2018). Bit by Bit. Social Research in the Digital Age, Princeton University Press, New Jersey.

UNECE (2013), Classification of Types of Big Data. [https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data](https://statswiki.unece.org/display/bigdata/Classification%2Bof%2BTypes%2Bof%2BBig%2BData)

Couper M. (2013), Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys, Survey Research Methods 7(3), pp. 145-156.

Haraldsen G. (2005), Using client side paradata as process quality indicators in web surveys. Exploratory workshop on Internet survey methodology: Towards concerted European research efforts, Dubrovnik, Croatia.

Struhl S. (2015), Practical Text Analytics, Kogan Page Limited, London.

Desamparados B., Reis F. and Domenech J. (2019), Forecasting tourist arrivals with online data: An application to the Valencian Community, NTTS 2019, Brussels, Belgium.