United Nations ECE/CES/2019/33

Economic and Social Council

Distr.: General
8 May 2019

English only

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Paris, 26-28 June 2019
Item 2 (a) of the provisional agenda
New data sources – accessibility and use
Session 2: Accessing new data sources

A statistical approach to Big Data

Note by Statistics Norway

Summary

This note presents reflections and suggestions made during Statistic Norway's work with a Big Data Strategy.

A statistical approach to Big Data should reflect the unique characteristics and purpose of statistics. There is a need for stable data sources, containing valid and reliable data about main social and economic processes. Moreover, we distinguish between a challenge-based approach, which addresses major challenges in present data collections and statistics, and an opportunity-oriented approach, focusing on the qualities and potentials of the different digital data sources. These two approaches can be combined and coordinated by a Center of Data Expertise.

Focus groups with producers of statistics revealed incomplete data as a major challenge; not only because of nonresponse in surveys but also in administrative registers. A major challenge with Big Data is that they do not easily fit into traditional data matrices. The note discusses alternative ways to establish data matrices, to produce statistics that do not rely on identified units and ways to combine surveys with Big Data.

This document is presented to the 2019 Conference of European Statisticians seminar on "New data sources – accessibility and use", session 1 "Accessing new data sources" for discussion.

I. Introduction

1. The term Big Data has become a catchword for a range of large, volatile, unstructured data sources. These data are already used for different purposes. Commercial actors are using Big Data to identify customer habits and tailor their sale strategies. Insurance companies, tax authorities and the police are using Big Data to reveal fraud and other criminal activity.

2. Statistical institutes are not commercial, and statistics is not about outliers but about averages and aggregates. Therefore, our approach and Big Data strategy should reflect the unique characteristics and purpose of statistics. Statistics can be defined by its activities (e.g. Dodge 2006 in Oxford Dictionary of Statistical Terms) or by its qualities, like in principle 11 to 15 in Eurostat’s Code of Practice (Eurostat 2018). For the purposes of this paper we would like to highlight the following characteristics:

- Statistics should be relevant. In its most general sense statistics should be of social relevance. In practice, however, the relevance of statistics will differ between different kinds of users like decision makers, mediators, businesses and private persons. Consequently, a wide range of statistical products and dissemination channels is needed.

- Statistics should be reliable and valid. These are the two basic quality requirements of statistics. They are affected both by the data source and by what data is collected. Generally, data validity, which is how well the collected data measure what they are meant to measure, is the most challenging part of data collection. In voluntary surveys, bias because of nonresponse is a main problem. We allow for uncertainty in statistics but strive to reduce it and decide the size of it.

- Comparability. Statistics is about comparing aggregates in time and between classified entities (like social groups or geographical areas). Breaks in time series or definitions cause problems 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.

3. A statistical approach to Big Data means looking for stable sources and reliable tools which can offer new or better insight into social and economic processes.

II. Big Data characteristics

4. The term Big Data refers first of all to the big volume of data, often specified 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. Robert Groves has suggested an alternative term “organic data” because it concerns data which reflect how a social organism is working (Groves 2013). This is an important point and we prefer to define Big Data as digital traces of human activities and attitudes (César Hidalgo referred to in (Few 2018)). Instead of asking people what they do and believe, these new data sources open a possibility to produce statistics based on the digital traces of actions and attitudes; be it process data or data about products and services. Using this definition as a starting point, we focus on how social and economic behavior are represented in digital sources. Looking at how technology also affects our behavior and attitudes will be the next step.

5. In addition to these common characteristics of Big Data, different kinds of digital sources have their own characteristics. We distinguish between three main kinds of Big Data:

(a) Transaction data which are registrations of economic or social exchange processes. These are the direct measurements of human actions but also the most protected kind of Big Data;

(b) Machine generated data, often called The Internet of Things, which are traces of behavior or products of behavior 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 1980s. They recorded what television channel people have chosen but not if they watched it;

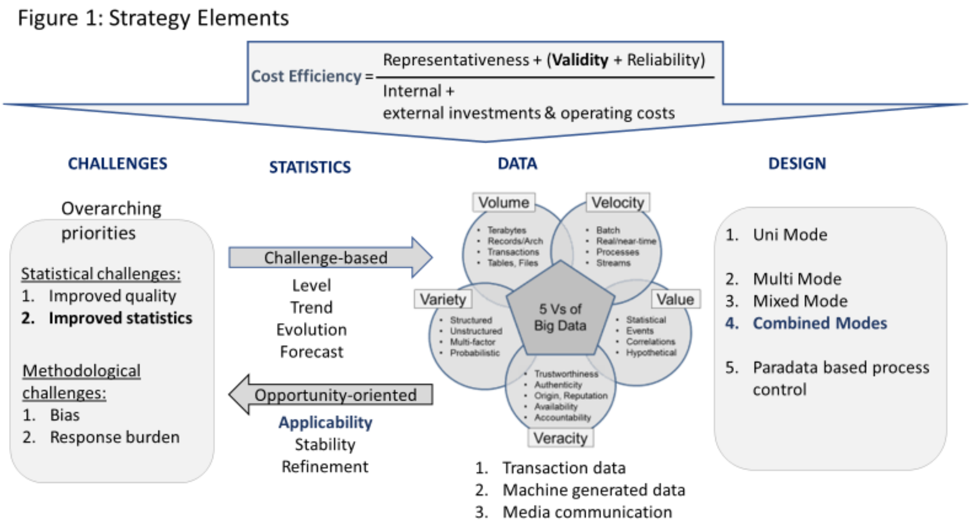
(c) Information or opinions communicated by broadcasting or in social media. Main focus has been on social media. Because this information and these opinions are initiated for a certain purpose, they may not give a representative picture of what people think and believe in (Couper 2013).

6. The potential and challenges of utilizing Big Data in statistics are both affected by characteristics common to this kind of data and by the particularities of different kinds of Big Data.

III. Strategies

7. Figure 1 aims to name elements which are considered for utilizing new data sources for statistical purposes. Cost efficiency defined as quality (representativeness, validity and reliability) over cost is the overriding criterion. In addition, the figure lists some elements and characteristics concerning challenges, statistics, Big Data and data collection designs.

Figure 1
Strategy Elements



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

9. The Danish Big Data Strategy is an example of a challenge-based approach. Its introduction states that “Statistics Denmark's Big Data Strategy will focus mainly on the application of data related to existing statistics (Denmark 2018). The term “opportunity oriented” is borrowed from CBS in the Netherlands. In 2016, they established an independent research and development environment, the Center for Big Data Statistics (CBDS) that looks at the possibilities for:

- Data that captures new phenomena and allows to produce statistics faster (“Real time statistics”)

- More detailed statistics, both as more specific measurements and as more detailed geographic disaggregations
- Data collection that involves lower or preferably no respondent burden ("Zero-footprint concept").

10. These three bullet points relate to some of the properties of the new data sources: that Big Data reflect behavior and attitudes and consequently may replace questions, that volume allows for more details, and that velocity allows for continuous updates.

11. We consider that a challenge-based and opportunity-oriented approach can be combined. Furthermore, the challenge-based approach should be organized as concrete projects initiated by statistical or methodological units while the opportunity-oriented approach should be run by an innovation group with methodology, research and computer scientists recruited according to an internship scheme. To maintain continuity, coordinate the work done, look for new opportunities and build competence. For this purpose, we envisage an overarching Centre of Data Expertise with cutting edge knowledge and ideas on how to extract valuable information from data.

IV. Priorities

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

13. The discussion revealed that incomplete data sources were considered to be a major challenge. It is well known 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 are health services and education offered on Internet and informal services such as 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.

14. By replacing survey questions with digital tracks, we reduce response burdens. This is particularly important in business surveys, where time is money, and in personal surveys concerning detailed questions about consumption, daily activities or travelling. A lower response burden in voluntary surveys may also contribute to a higher response rate.

15. There is also a demand for more direct measurements of economic transactions, for example of the price of goods purchased instead of commodity prices from one source, which is adjusted for shopping patterns gathered among different kinds of customers from another source.

16. Voluminous datasets may offer interesting opportunities for more detailed health statistics, immigration statistics, innovation statistics and others. More frequent statistics are required in some areas, in particular within 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 that what is the case.

17. 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 tracks that can be used to analyse and improve the recruitment and response processes, preferably along the way during the data collections (what is commonly called a responsive approach). In this way we can use Big Data tools to study our own behaviour and improve our own work processes.

18. Based on these results and reflections, four challenge-based issues are listed in the right upper part of figure 2 which may be addressed by Big Data. The lower part provides suggestions for the priorities of an opportunity-oriented approach. Reasons for these priorities will be explained in more detail in the next section.

Figure 2
Priorities

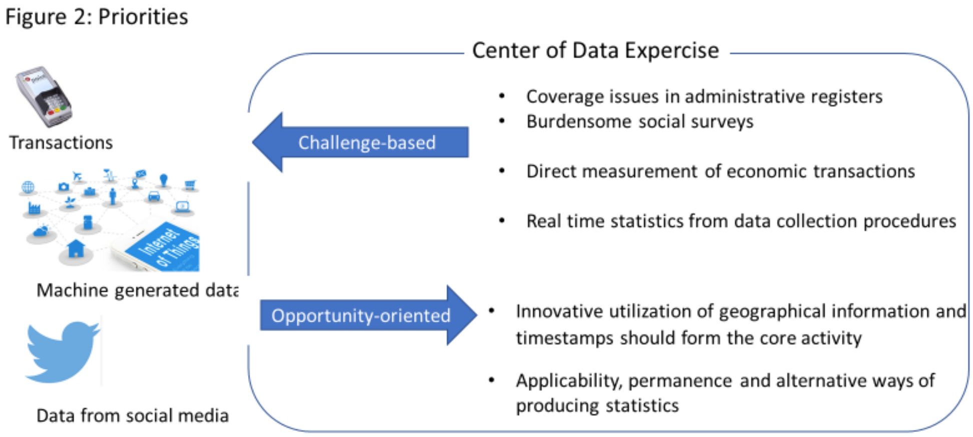


Figure 2: Priorities
Transactions
Machine generated data
Data from social media
Center of Data Expercise
Challenge-based
Opportunity-oriented
• Coverage issues in administrative registers
• Burdensome social surveys
• Direct measurement of economic transactions
• Real time statistics from data collection procedures
• Innovative utilization of geographical information and timestamps should form the core activity
• Applicability, permanence and alternative ways of producing statistics

V. Alternatives

19. Traditional data collections are designed to collect data relevant to a certain statistic from a population or sample of identifiable units. The delivery is a data matrix with unit identifications at the left of each line and the different data variable along the following columns. This matrix will be the basis for statistical analysis. Data capture from Big Data sources is different from this in two ways.

20. In designed data collections we ask a set of questions which are intended to collect certain kinds of information first and collect answers next. When capturing 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 analytical approach, similar to the approach taken by media scientists when they study patterns and trends in media content. Content analyses are 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).

21. To capture Big Data is more complicated because of the volume and the amount 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 challenge, however, is that the data captured seldom fit into a traditional data matrix. Persons, enterprises or other units may not identifiable, and there are few clues to how the data should be interpreted. 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

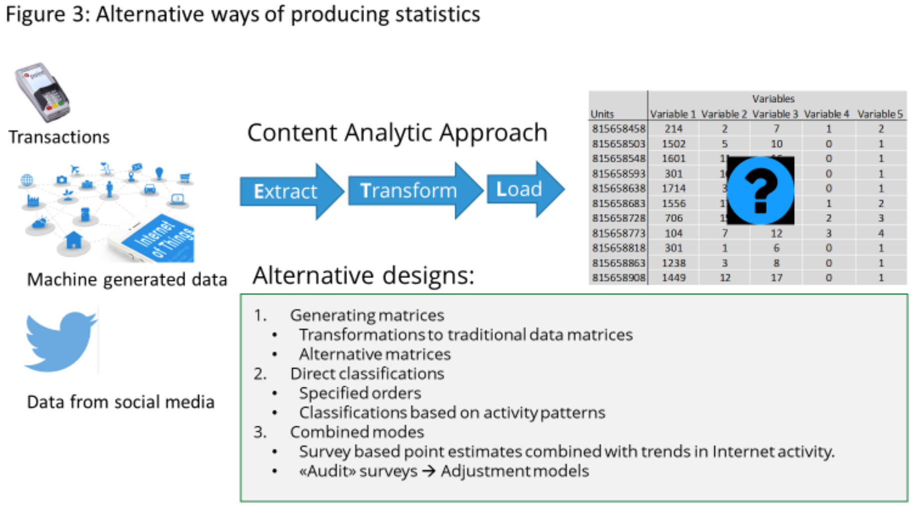


Figure 3: Alternative ways of producing statistics

	Variables				
Units	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
815658458	214	2	7	1	2
815658503	1502	5	10	0	1
815658548	1601	[illegible]	[illegible]	0	1
815658593	301	[illegible]	[illegible]	0	1
815658638	1714	[illegible]	[illegible]	0	1
815658683	1556	[illegible]	[illegible]	1	2
815658728	706	[illegible]	[illegible]	2	3
815658773	104	7	12	3	4
815658818	301	1	6	0	1
815658863	1238	3	8	0	1
815658908	1449	12	17	0	1

Alternative designs:

1. Generating matrices
   - Transformations to traditional data matrices
   - Alternative matrices
2. Direct classifications
   - Specified orders
   - Classifications based on activity patterns
3. Combined modes
   - Survey based point estimates combined with trends in Internet activity.
   - «Audit» surveys → Adjustment models

22. 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 Statistic Norway’s effort to link receipts from grocery stores with account details from banks. The stores keep track of what goods have been sold and how much has been paid for each item, but they do not know who the customers were.

23. Most Norwegians purchase their groceries with debit cards and the grocery market is dominated by three large grocery chains. When paid for by card the bank will know the name of the shop, the total paid and the account from which the amount should be deducted. The owner of the account is identified by his or her personal ID. The ID contains information about age and sex and can be linked to other kinds of background information in different statistical registers. Details about what was paid for, however, do not follow the deduction information.

24. In 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 rates in these surveys are unacceptably low. If we can link 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 at the bank transaction in combination with the name of the store. From this a traditional matrix with identifiable customers and goods bought can be constructed.

25. In the grocery example we need the personal ID to study grocery expenditures in different parts of the population. Likewise, in economic statistics we use enterprise and establishment IDs to compare different kinds of businesses. These are the two most common identifiers. But there are alternatives and some of these are both easier to find in Big Data and interesting to use in statistics. One example is a study of innovative small businesses conducted by Statistics Netherlands. In this study geographical areas are used as units and innovativeness measured by web scraping data from business websites (van der Doef 2018). Together they form an alternative matrix.

26. It should be noted that timestamps and geographical coordinates play key roles in these two examples.

27. ID number is only a classification tool in statistics. As soon as the units are classified, data can be anonymised. A different approach when identifiers are missing is therefore to look for alternative classification methods. Identifiers may exist but be missing 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 order statistics from a different institution. 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.

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

- if you have small children (because you regularly go to kindergartens)
- if you are working full time (because you normally go to and leave the same working place each weekday)
- what kind of branch you are working in (because we know what is produced at different sites).

29. 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 is 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). In the design box in Figure 1 a third alternative is suggested, namely combined modes. Combined modes are when different modes are used for different purposes. Under the Statistics heading in Figure 1, four kinds of statistics are identified: level, trend, evolution and forecast. These are often combined, e.g. estimating average prices (level), how prices have changed (trend) and will change in the future (forecast). 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 customers shopping across the border at a certain time. The result is then compared with mobile phone activity abroad, 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 changes (e.g. Desamparados 2019).

30. The last option listed 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 examining the quality of services offered or products 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 of adjustments. Considering that Big Data often is presented as a replacement that will take over for surveys, this is truly a paradox.

References

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

Denmark, S. (2018). “Big Data Strategy 2018-2020”.

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

Eurostat (2018). European Statistics Code of Practice. Luxembourg, European Union.

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

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

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

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

Struijs, P., De Broe, S. (2018). Big Data Strategies for Official Statistics. DGINS 2018. Bucharest, Romania.

van der Doef, S., Daas, P., Windmeijer, D. (2018). Identifying Innovative Companies from their Website. BigSurv18. Barcelona, Spain.