United Nations ECE/CES/2014/32

Economic and Social Council

Distr.: General
19 March 2014

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-second plenary session
Paris, 9-11 April 2014
Item 3 of the provisional agenda
What is the value of official statistics and how do we communicate that value?

Big data – an opportunity or a threat to official statistics?

Prepared by Eurostat

Summary

This paper summarises the main questions at stake regarding the issue of Big Data and official statistics, and provides suggestions on how to tackle the problems. The discussion on Big Data is now on the agenda of many conferences and meetings in official statistics. At the same time the public debate of its advantages and disadvantages is ongoing. What can we say about the role of official statistics in a world of continuously increasing new data sources? Does this put any pressure to change the way in which official statistics are produced? Is the deluge of new statistical data instantly available from various sources going to substitute the production of official figures?

The paper is presented for discussion to the Conference of European Statisticians seminar on "What is the value of official statistics and how do we communicate that value?"

I. The role of official statistics and statisticians

1. There is certain agreement between users and in public opinion that the term "Official Statistics" refers to figures that must be generated:

(a) with the purpose to serve the whole spectrum of the society;

(b) based on quality criteria and best practices;

(c) by statisticians with assured professional independence and objectivity.

2. Statistics produced following this approach are a key pillar of democratic societies, providing a quantitative assessment of governments' policies and allowing for comparisons between countries, regions or effects of alternative actions. These ideas have been generally recognised and accepted, and therefore implemented in the statistical laws of many countries around the world. Professionals working in the production of official statistics may be seen as some sort of keepers of the truth. Their job is usually carried out as civil servants in institutional bodies, often with a strong background in statistical science. This organization of work is accepted as a guarantee that ensures the confidence of the society in the reliability of the statistics produced.

3. A crucial issue to be considered is the quality of official statistics. Traditional definition of quality as statistical accuracy based on sampling theory has been complemented over the years by more user-oriented quality criteria such as relevance, timeliness, coherence or integrity (Eurostat, 2011).

4. On the other hand, the Big Data industry is rising: the huge volume of digital information derived from all types of human activities is being increasingly exploited to produce statistical figures. These figures often make use of data from private institutions or companies. Leaving aside the current public debate on whether companies which collect the data should own the data and could use them for another purpose without consent, these new statistical figures may be seen as competitors of traditional official statistics.

A. A risk that should be avoided

5. The production of statistics is a specialised task requiring certain background and knowledge. This makes it difficult to define the actual meaning of the word "official" in the context of statistical quality. The reason is that while some of the requirements for achieving high quality can be assessed following objective criteria – for example, using commonly agreed quality criteria and best practices – others are based on ethics and professionalism of statisticians.

6. As a consequence, sometimes official figures published by well-known statistical agencies may later need to be corrected or revised. To this end, bodies with the purpose of monitoring and ensuring the quality of official statistics have been created, such as the European Statistical Governance Advisory Board (ESGAB) and others.

7. Considering the heavy procedures in place in the official statistical system for quality monitoring – including actions of national statistical offices and other bodies with auxiliary, coordination, monitoring or advisory roles – it seems improbable that statistics derived from Big Data by businesses could acquire the same qualifications. While this statement is true, still there are some risks that can lead to lower confidence in the current system of official statistics.

8. One of the dimensions of quality in statistics is timeliness which is of growing interest in the globalized world. Some phases in the statistical process involve time-consuming tasks frequently carried out in an iterative way. There is clearly a trade-off between timeliness and accuracy until the final figures are delivered. Statistics derived from Big Data may have a competitive advantage in this regard. One of the reasons of better timeliness is that the data collection phase of the generic statistical process – including all its sub-processes – is commonly substituted in Big Data sources by automatic and instantaneous data availability.

9. This feature may influence in particular short-term indicators. It is possible that indicators based on Big Data sources could follow the evolution of some of the key short-term variables with some accuracy. Politicians and other data users could actually decide to use the first available indicator, especially if problems in accuracy are rewarded by notably shorter delays. This could make some of the official short-term statistics irrelevant and gradually hamper the confidence in the current statistical system. However, statistics derived from Big Data sources do not provide any guarantee of high quality, such as professional independence and regular availability. Losing official data sources could in the end result in a loss of transparency with impacts on democracy.

B. Big Data as a big opportunity for official statistics

10. Is there any way to prevent the possible harm to the system of official statistics from happening? The answer could be to explore the possibilities of using Big Data sources in official statistics. Several national statistical offices, Eurostat and other bodies have started to draft a road map for doing so. Nevertheless, there are some interesting issues to consider with the purpose of improving competitiveness of official statistics, because Big Data may help improve with timeliness of statistics.

11. Big Data can be used for the production of official statistics in different ways: (i) replacing statistical sources completely, based on common definitions, classifications, etc., (ii) partially replacing statistical sources, completing the information by means of record linkage, matching or other procedures, and (iii) providing completely new statistical figures that may complement the available statistical information. The two first ways may result – in theory – on significant reductions of costs and respondent burden, but they would imply new tasks of translating, linking or harmonizing different structures of data to statistical classifications, definitions, etc. These tasks are not necessary when completely new statistics are created and produced. Big Data, thus, may not be able to totally or partially replace statistical sources in the short term and following such a strategy might be too expensive in terms of time and other resources. Businesses producing statistics based on Big Data follow the third way, not having to cope with these problems.

12. Big Data sources offer huge volumes of data which require storage and processing that exceed the capacity of traditional statistical production tools. Using Big Data would necessitate moving away from exclusive dependence on statistical methods that cannot handle huge volumes of information. Instead, a more diverse set of tools should be adopted. This can be addressed through the use of data mining and machine learning algorithms having the required computational efficiency (Bondi, 2000). Besides, due to the enormous heterogeneity of Big Data sources as regards formats, contents, storage, etc., methods to produce statistical information should be developed *ad hoc* for each case and the Generic Statistical Business Process Model might not be applicable. In the beginning, the Big Data source should be explored using data mining procedures to learn about the unknown data structure and decide on the possible outcomes, and how to combine these data with traditional statistical production procedures.

13. A third consideration refers to the representativeness and coverage of the statistics produced. The use of probabilistic sampling in traditional statistics provides a theoretical framework that ensures confidence on the figures based on sampling errors. Most of Big Data available cannot be adapted to this framework and other procedures should be developed. This seems to be an important weakness and efforts should be focused on it. Meanwhile, the experience of successful uses of Big Data could be explored so that countries could learn from others' experience.

14. Some of the known statistics derived from Big Data sources measure the evolution of variables that are a proxy to typical key economic and social variables. A well-established principle in statistical practice is that changes (over time or space) can be estimated more reliably than absolute figures. Maybe the first attempts to use Big Data should address producing indicators measuring change. This would allow using a similar approach for judging the reliability of the figures: assessing the performance in terms of similarity or correlation to other figures measuring the same or an analogous phenomenon. It makes sense from a data mining perspective, where the alternative to fitting a statistical model is using an algorithm reflecting the real world. When many different statistical figures are produced from different and independent Big Data sources following these principles, the coherence and agreement among them may be an argument to support the validity and representativeness of the whole set.

15. Consequently, there is an opportunity to establish a strategy to exploit Big Data to produce competitive statistics with acceptable levels of quality. It can be briefly summarised in the following points:

(a) Start producing short term indicators on economic and social phenomena without transferring Big Data structures into statistical structures but using their own. An example would be to build a monthly indicator of the evolution of the household budget based on a Big Data source using its own classification of goods and services, and harmonize with a classical statistical classification at the highest level of the hierarchy at the maximum;

(b) Use other statistical figures available for the same or related phenomenon (timely data, related definitions, common or similar geographic frames…) as a framework for comparing and studying representativeness and coverage. In the previous example, the evolution of the produced households' budget indicator could be evaluated through its comparison to the evolution of the annual Households Budget Survey in the same or similar geographical areas at the global level and –if possible– at other levels of the hierarchical classification;

(c) Instead of the first phases in the Generic Statistical Business Process Model, perform an exploratory and research step using data mining and/or machine learning procedures to start using a new Big Data source. The exploratory step on the transactions file of a department store, for example, could be performed using association rules analysis (Zaki, 2000), which would provide the shopping baskets more commonly sold together. Of course, a different Big Data file would need a different first step to explore its contents;

(d) Continue with a combination of data mining, machine learning and statistical procedures to produce the statistical figures. In the example above, the confidence[1] and the support[2] of the association rules (Cios et al., 2010), jointly with the shopping baskets

[1] The confidence of an association rule is defined as the ratio of the number of transactions containing the items of both sides of the rule to the number of transactions containing the left side items.

[2] The support of a rule indicates the frequency (probability) of the entire rule with respect to the whole set of transactions.

commonly sold together, can be studied to build appropriate composite index numbers to measure the evolution of the households' budget.

16. The short-term indicators produced from Big Data following the above ideas can be a good complement to traditional statistics. They would provide the means to roughly update the figures until the next structural survey or census is conducted. Gaining experience in producing these indicators could help official statisticians to learn about possible new methods of data processing and start redesigning the official statistical systems to build a new paradigm.

17. Also in this direction, the role of some of the statistical infrastructures that hinder development must be reconsidered. In particular, statistical classifications require huge efforts to be built, are not easy to maintain up to date with real developments and thus need to be continuously revised. This increases the difficulties in the data collection phase – perhaps the problem could be addressed in another way. For example, why build detailed classifications of economic activities? Data mining methods for clustering or text mining based on the description of activities by companies and/or the input and output products could provide more realistic categories. The saving of costs and the reduction of delays would compensate for the possible loss of quality.

18. A final remark is that prior to engaging in a complex process for producing statistics from a Big Data source, a careful analysis of the potential gains should be made, taking into account that the advantages may balance a possible decrease of accuracy or quality in general.

19. The advances in the proposed strategy may signify a big step to achieving new official statistics systems able to maintain the confidence of society, in a less expensive way while reducing respondent burden.

References

- Bondi, A. B. (2000), "Characteristics of scalability and their impact on performance", Proceedings of the 2nd international workshop on Software and performance, Ottawa, Ontario, Canada, ISBN 1-58113-195-X.
- Cios, K. J., Pedrycz, W., Swiniarski, R. W. and Kurgan, L. A. (2010), Data Mining: A Knowledge Discovery Approach, Springer P. C., Incorp., 1st ed.
- Eurostat (2011), Quality Assurance Framework of the European Statistical System, version 1.1, Eurostat.
- Zaki, M. J. (2000), "Scalable algorithms for association mining", IEEE Transactions on Knowledge and Data Engineering, vol. 12 (3), pp.372-390.