United Nations ECE/CES/GE.41/2016/10

Economic and Social Council

Distr.: General
19 July 2016

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Eighteenth Meeting
Geneva, 28 - 30 September 2016
Item 7 of the provisional agenda
Possible uses of new data sources (e.g. "Big Data") for censuses

Is it possible to use Big Data in the 2020 Census Round?

Note by the Central Statistical Office of Poland (CSO)

Summary

The potential of new Big Data sources and technologies is very important for official statistics, but can Big Data have a significant impact on the next census round? The majority of States would like to carry out the Census 2020 at a lower cost than the previous 2010 Census and maintain high quality results. The large volume of data is not new for many statistical offices because in the previous Census 2010 many countries have used administrative data. However, Big Data is a very complex and relatively new subject.

Big Data sources have great potential in the context of censuses as crucial statistical surveys. However, the immediate perspective of 2020 census round is so close that there is a high probability that the Big Data source will not be properly investigated at the time, described and practically used for statistical research on a smaller scale. There are many methodological, legal, quality and IT challenges for the use of Big Data within official statistics frameworks, and continuing work in this area is a high priority. Therefore, further work on Big Data is still needed, and in the case of positive and promising results the use of these sources in the next (i.e. after 2020 round) censuses rounds could be recommended.

GE.16-12448(E)

1612448

I. Introduction

1. The phenomenon of Big Data is becoming more and more popular in the world. It is a global trend of continuous searching for new things, a constant need to develop and improve the world around us. This phenomenon is most visible in business - private sector, which primary goal is profit. It refers also to the public sector. Public administration, although committed to other purposes, such as optimizing public costs, improving the functioning of public institutions or orientation of the citizen, also sees a potential in using Big Data. Also official statistics sees in Big Data a great opportunity. By combining new data sources and new experts the statistical world is able to analyze so far unexplored phenomenon and draw conclusions about what scientists 10 years ago never dreamed.

II. Using Big Data in Official Statistics

2. A typical data sources can open the statistics for a wide range of possibilities. Currently official statistics, as well as all kinds of indicators and metrics, are based on data from state registers and information obtained from entrepreneurs, respondents or acquired on the basis of observations of interviewers and experts. Registers, observations and reports submitted by the companies provide hard data, interviews with respondents are the source of opinions, moods and subjective data.

3. However, the world is continually changing and there are new phenomena which also require describing with statistical processes. Therefore it cannot be limited only to the known data sources, you need to constantly seek new paths and solutions. This will also help in a new way to look at the current working methods.

4. An example would be the development of ICT survey in Poland. Currently to study such phenomena, questionnaires are used, among others. Respondents answer to questions related to the use of their computers and the Internet. The use of Big Data in this particular area means e.g. creating a set of robots scanning Web sites and studying their content based on previously entered patterns. The results give answers the questions about using ICT in enterprises, among others, type of website, owned online store, the type of available data logging capabilities, etc.

5. Big Data is also widely used in the area of labour market. By proper analysis of websites, social networks and news sites e.g. by using certain key words or data mining algorithms concerning jobs by occupation, region (province), ownership sector, statistics can complement the data on the labour market, especially in the context of the demand for labour , employee turnover and strikes.

6. On the other hand, data from mobile phone operators help in the study of the phenomenon of population movements during the day and night, which is a perfect base for a more detailed analysis of the causes of these migrations.

7. Traffic, including quantitative flow and type of vehicle, can be monitored by cameras and sensors placed at selected locations road. As a result, e.g. in surveys conducted at border crossings you can minimize or at least reduce job interviewers counting vehicles.

8. The use of Big Data, not only would complement the data available to the statistics, but in the distant future would enable the replacement of part of the currently existing conventional surveys. As a result, it would enable to reduce the involvement of citizens in filling in questionnaires and statistical interviewers in collecting these questionnaires, while maintaining the existing high-quality data.

III. Chosen international works

9. Works on the Big Data implementation to official statistics at the international level have lasted since the last few years. Both institutions at the global and European level are involved, as well as representatives of business, academia. Among several initiatives, one worth mentioning is the UNECE initiative. UNECE created a project with a collaborative approach, bringing together over 70 experts from national and international statistical organizations around the world to identify and tackle the main challenges of using Big Data sources for official statistics. The project ran from January to December 2014 and had three main objectives:

- To identify, examine and provide guidance for statistical organizations on the main strategic and methodological issues that Big Data poses for the official statistics industry;
- To demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts;
- To facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

10. The project consisted of several task teams responsible for relating issues:

- Partnerships Task Team;
- Privacy Task Team;
- Quality Task Team;
- Sandbox Task Team. Works of this team were continued also in 2015.

11. The result of the project was a series of documents relating to various aspects, which task teams were engaged in, as well as creating Big Data Inventory.

12. Another important initiative is the ESSnet Big Data Project - part of the Big Data Action Plan and Roadmap which goal is to integrate BDAR into the ESS Vision 2020 portfolio.

13. This project has a dual purpose. The first is running pilot projects exploring the potential of selected Big Data sources for producing or contributing to the production of official statistics. The second is composed of horizontal topics, such as methodology, quality, IT infrastructure and metadata, which lay down the conditions for future use of these data sources within the European Statistical System.

14. The overall objective of the project is to prepare the ESS for integration of Big Data sources into the production of official statistics. The award criteria of the FPA mention that the project must focus on running pilot projects exploring the potential of selected big data sources for producing or contributing to the production of official statistics. Aim of these pilots is to obtain hands-on experience in the use of big data for official statistics.

15. The specific objectives of the ESSnet Big Data are:

- Pilots for generating statistics from big data sources at ESS level;
- Analysis of the output portfolio of Big Data sources;
- Development/Review of methodological and quality frameworks for Big Data sources in official statistics;

- Identification, definition and implementation of IT infrastructures for Big Data processing;
- Access to Big Data sources, identification and preparation of non-legal and legal conditions for access and use of Big Bata within the ESS;
- Exchange of information with stakeholders within the statistical system and the research community.

IV. Censuses

16. The trend of implementing Big Data to official statistics tends to increase and the potential of Big Data sources is huge. Here comes a natural question: if proven that Big Data can be successfully included to the official statistics on the example of surveys, is it possible to use it already in the 2020 census round?

17. To answer this question you first need to consider the specifics of censuses. The census provides the most detailed information on population; its territorial location; the demographic, social and professional structure; as well as socio-economic characteristics of households and families and their resources; housing conditions at all levels of the territorial division of the country: a national, regional and local. Particular attention is given to the acquisition of knowledge on changes in demographic and social processes, especially due to increased migration. The results of censuses are used directly for the needs of official statistics as a base to build sampling frames for next surveys conducted on a sample of households. The most important include obtain information about the issues that were covered by the census in the previous round. The continuation is necessary to conduct comparative analyzes of phenomena in time and to describe the changes that have occurred in the process of demographic, social and economic in terms of: population, the state of housing and buildings; households and families in conjunction with the housing conditions. This approach allows to keep time series and comparability with the results of previous censuses.

18. The credibility of the census depends on the fulfillment of the following conditions:

- Universality - Censuses conducted across the country without the omission of any region or area;
- The collected data must refer to a specific time (a specific day and a specific time);
- Directness - information should be obtained directly from the respondents or from referenced sources like good quality administrative registers;
- Conducting only for statistical purposes, which means that there is a secret census and the data obtained cannot be used for other purposes.

19. It appears, therefore, that the census is a far more significant survey than the other ones. The frequency of its conduct, often amounting to 10 years means that once collected data must be a source of knowledge in some areas up to a decade. For this reason census results are often crucial for creating economic and demographic strategies. The census must therefore be conducted in a fair, accurate and professional way. To meet these conditions it is necessary to have an appropriate methodology, tools, knowledge and competence of statistics staff.

V. Obstacles

20. Due to the fact that Big Data is a relatively new phenomenon, many countries have not yet developed clear guidelines for collection, processing and sharing such data. Because of these limitations, many aspects of this phenomenon have not been tested in a practical manner. Even in the case of implementation of various kinds of practical pilot projects using Big Data sources because of the short period of time we have not yet a reliable, robust method of processing and data analysis. The quality of products developed on the basis of Big Data, of course, is increasing, but in the case of such a unique survey as the census no one can afford mistakes or uncertain quality.

21. Another problem arising both on the international and national level is the lack of experts - so-called data scientists. The desire to reach for a variety of data generated by companies or by people, e.g. on social networking sites; data generated by public institutions; data generated by all kinds of devices, machines became a stimulus for growth in demand for professionals with exceptional skills. Recently, the combination of knowledge, competencies and skills in areas such as information technology, processes, business optimization, mathematics and statistics allowed to evolve a modern data science, and specialists - ranging from “traditional” Data/Business Analyst after those of processing larger amounts of data: Data Engineer/BI Engineer/ Data Architect who can expand their knowledge, skills and competences in other fields of science, business, as well as new technologies, sources and channels of information transmission came to be regarded as a Data Scientist/ Big Data Engineer.

22. Data Scientist is an example of a modern specialist. Unlike a specialist in the usual sense, who was an expert in one area, this new one being an expert in its core business, e.g. computer science widens his/hers knowledge of quite opposite areas e.g. psychology. The combination of those seemingly totally unrelated skills, gives a very fruitful mix. It allows to analyze unstructured, dynamic data in an equally dynamic and unusual way. This is the reason why it is extremely difficult to find people who meet these criteria. There are, of course, scientists possessing skills, but to be able to get benefits Big Data - something more is needed. Therefore, one of the key issues is to provide a range of IT training, analytical, methodological, etc. How, then, while not having appropriately qualified people, carry out such a serious survey as a census? It would be a very risky step and finally could lead to errors.

23. One of the major obstacles is the lack of sufficient legal basis for collection, analysis and storage of Big Data. A scale of the problem is very diverse in the world. Each state has its own regulations, which make it more difficult for some, than others, to reach for unusual data not included in the existing legal system. Then there are sector confidentialities, which often prevent owners to share their data for statistical purposes. Although carrying out the census using Big Data would certainly be an interesting project, however, you have to consider if it possible to do it in 2020 or during the next census round. Apart from the above-described problems with quality, lack of specialists or methodology, you should think whether in 2020 the data, without which the inventory cannot take place will be available at all.

VI. Conclusion

24. To sum up, we can say that the use of Big Data, although it has many advantages, however, is also connected with certain restrictions, which are largely critical and sometimes restrain further work.

25. Unfortunately, the implementation of Big Data technology in the public sector, including statistics, is not easy. The public sector has a much more difficult task in this context than the private one. In business, which is focused on the issue of profit, customer data protection is associated largely with the image of the company. In the Internet era, when unflattering opinions can disperse like wildfire, no company can afford to lose reputation. This, combined with all kinds of legal safeguards, guarantees data protection, but does not paralyze development of the company. Hence, with the appropriate safeguarding customer data solutions, Big Data technologies can be used in a variety of ways.

26. The situation is different in the public sector. Administration is not focused on profit but on the service to the citizens. In this context, state institutions must be flawless bodies, because they represent the dignity of state. This means that the need of data security is even higher than in the case of business. The law is a guard of security, and this law can be very restrictive.

27. Therefore, the official statistics cannot afford to rush and hasty action. It must uphold the quality and integrity of all data which is collected, in particular - data from censuses. It does not mean stopping the search for new paths. Big Data is a good clue, just a little too fresh. All initiatives, which were mentioned above, will soon give the answer how great potential for statistics is Big Data. After a more careful and exact analysis it will gradually turn out which, at first smaller and less complicated surveys, can be supplemented by Big Data. With time, the circle of surveys in which Big Data will be used will be expanded. There is a chance that in a few years we will be witnesses of the replacement of current collecting data methods by new opportunities. Taking the above into consideration, the use of Big Data will not be possible in the 2020 census round but it will be possible after the year 2024.