Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Topic: Data: 2021 Census, administrative data, geospatial data, big data and other alternative data.

Title of the presentation: Use of administrative data and alternative data for census when applying modern technologies

Prepared by Janusz Dygaszewicz, Director, ICT Systems, Geostatistics and Census Department, Statistics Poland

I. Introduction

1. The population and housing census is a statistical survey of exceptional scale and importance. It is a survey incomparable to any other statistical survey, because it concerns the entire population of the country, and its purpose is to collect and develop complete information on the condition and socio-demographic structure of the population and housing resources of the country. It is implemented every 10 years. During this time, both the census methodology and the methodology of data collection from respondents are being developed. The market for technological tools and ICT systems, which significantly affect the organisation of data collection methods and the participation of respondents in the census, is changing rapidly. The growing circle of possibilities to address information messages to respondents, including digital tools, helps to shape pro-social and "pro-statistical" attitudes (affecting the image of official statistics).

2. Taking into account the aforementioned arguments, it can be assumed that in the population and housing censuses over the years there are constantly changes aimed at improving the quality of censuses. Only the principle of preserving the comparability of results and continuing the creation of time series remains unchanged.

3. Looking back at the past census, Poland was one of the first countries in the world which prepared a totally innovative method consisting of using several of the most modern techniques for collecting census data simultaneously. Data from 28 administrative registers and 3 non-administrative systems were effectively integrated. Paper questionnaires were completely eliminated, and were replaced by ICT solutions. The use of GIS technology helped to conduct the census preparatory work and an on-going census process monitoring and give possibility to compile and present census results based on multi-dimensional spatial analyses. Apart from the use of IT systems and registers of public administration, various data collection methods were applied, based on functioning of three electronic channels simultaneously:

(a) CAII/CAWI (Computer Assisted Internet/Web Interview)
(b) CATI (Computer Assisted Telephone Interview)
(c) CAPI (Computer Assisted Personal Interview).

4. Information and communication technologies progress, creation of new data sources and related enormous amounts of data, construction of new registers and improvement of existing ones by public

administration as well as other non-public entities create conditions for use in censuses new data sources and continuation of using modern technologies data collection, initiated in the census 2011.

II. CAxI Methods

5. The use of CAxI channels for data collection allowed to reduce the number of census enumerators, which resulted in reducing the costs of the census. At the stage of collecting data, the cheapest method is the CAII channel, because no costs are incurred. In CATI and CAPI channels, costs related to salary for enumerators are borne, but rates are definitely higher for enumerators performing direct interviews (CAPI) due to travel costs to respondents.

6. The lessons learned from the census in 2011 lead to attempts to implement a full census 2021 instead of a representative survey as it was in 2011, while using data from registers as a census database. It will therefore be an evolutionary improvement of the combined census. The results obtained will be directly processed without the need to make estimates.

7. Replacing the representative survey with the full survey on all population will be the most important change. However, in order for it to bring the expected quality effect, the terms of the census must be reformulated. They concern the census form and the issue of census promotion. Other changes will result from the development of modern information technologies.

8. In order to encourage respondents to make a self-enumeration, it is necessary to create the best conditions for them, that is, to develop a clear, easy to use, short census electronic questionnaire and a friendly IT environment. The application that services the form should be available on any electronic equipment regardless of whether it is a tablet or a smartphone or another device. Also software that supports census systems should take into account the need to communicate with the software and devices used by the respondents.

9. In order to enable persons who do not have technical and material conditions to fulfill their self-census obligation, the commune office provided (during their work hours) free access to rooms equipped with computer with software installed sufficient to conduct an online self-census.

10. At the request of a person, directed to the commune head, while conducting an online self-census in a designated room, the necessary assistance in operating the interactive self-census application was provided.

11. What is important, additionally it will be possible to enumerate the respondent in the CATI channel "at the request", which was impossible in the 2011 census.

12. Popularization of the census should be based on modern tools and communication channels and should take into account the psychological side of information transmission and encourage not only participation in the census, but also in the most desirable form of this participation, i.e. through a self-enumeration. It should be remembered that the information message must be appropriately adapted to the respondent's profile and skillfully distributed over time and intensified as the date of the census approaches. During the census, promotional messages will be adapted to the course of the census. Employees of statistics do not have such a vocational preparation to meet the requirements of modern popularization, so it is planned to entrust the task of disseminating information about the census to a professional company that will be able to guarantee the best possible effect.

13. CAxI techniques used to collect data have a huge impact on the quality of results. Regardless of the channel in which the census is made, the friendliness and communication of the e-form increases the respondent's interest and the willingness to participate in the census. The unambiguity of questions avoids mistakes resulting from misunderstanding of questions. Clear answers to clearly defined questions also have influence on the amount of errors corrected as a result of data processing, as well as on the rate at which results are obtained, developed and made available. The preparation of the census results is an

activity summarizing the entire process of census implementation. Therefore, it should be done in a perfect and fast manner. The use of modern methods of data collection significantly speeds up their processing and development. Dissemination of census data in the form of micro-aggregates, aggregates and result sets should be implemented in an innovative way - by granting access to the analytical database. People who, for various reasons, use the results of censuses, are more likely to be interested in them the more their curiosity is stimulated through the friendly preparation of an e-census form. The information obtained by CAxI methods allow for conducting a variety of analyzes (including spatial analysis) and enables recipients to create individual reports. These reports will also be the better, the more precise and unambiguous the wording used in the census form. Thanks to the use of CAxI data collection channels, it will be possible to reduce more than five times the number of enumerators to carry out direct interviews in relation to the number of enumerators, which would be needed in the case of a traditional census.

14. In order to maintain statistical confidentiality, ensuring data protection alternative authentication methods in CAWI based on:

(a) national node (including Trusted Profile and bank IT systems),
(b) Personal Identification Number (PIN) associated with personal data contained in administrative registers
(c) unique 10-character flat code - assigned to each flat included in the census list

are examined and will be used in 2021 census.

III. Use of administrative data

15. For the purposes of censuses, data from multiple sources, including administrative and non-administrative sources, are collected. Registries and database systems are characterized by a wide variety and complexity resulting from the fact that they are created for different purposes and are managed by different data owners. Therefore, they also vary according to the used standards of the storage, accuracy or recording methods. The lack of uniformity exists not only between the registers but also in the data inside the registers.

16. The quality of the sources and data contained has an effect on the performance of the survey. Therefore, adequate quality is a prerequisite (although not the only one) to obtain correct census results. Thus, when using administrative and non-administrative sources in research, the key elements are to identify and understand problems encountered in the data, and then unify and correct them.

17. The selection of the sources for the National Census 2021 was based on the experience of the previous census and the data obtained annually for statistical surveys. Collected resources have been systematically analyzed. There is no single indicator that assesses a register in terms of subject, object content and integration with other registers. Datasets with a minimum quality criterion are downloaded and the process of their correction and enrichment are followed. Cooperation with administrative data owners seems to be of key importance. In Poland it is based on:

(d) long-term;
(e) effective;
(f) transparent;
(g) based on mutual trust;
(h) supported by a legal basis (Act on Statistics, Census Act).

18. In connection with the implementation of activities aimed at increasing the use of data from administrative sources, the list of people will be created only on data from administrative sources

IV. Geospatial data

19. Census information at accurate and standard geographic levels is essential to facilitate comparative analysis and achieve better quality of geospatial statistical data production. In the context of census it is crucial to have georeferenced data at the level of x,y coordinates.

20. The location of a given census units should be specified by means of GPS coordinates and descriptive attributes, especially taking into account the register identifiers of territorial units and address features (including the identifiers of administrative division units, locality identifiers and street identifiers).

21. For the purpose of enumeration conduction and results dissemination, the address point coordinates should be determined with accuracy of several meters (no more than 7 m).

22. With the aim of facilitating the management of spatial data concerning the location of census units, the coordinate system adopted should be homogenous for the whole country.

23. The introduction of x,y coordinates and address points in census data enabled changing of the previous system of spatial identification and shifting from area assignment (census districts) to point assignment. The change of the assignment mode allowed for more flexible grouping and presenting data collected by public statistics in statistical units smaller than the commune, i.e. in statistical regions, census enumeration areas or even in very small areas, such as a kilometer grid with a cell size of 1 km2. It also facilitated the creation of a spatially-oriented micro database, enabling the conduction of geo-statistical analyses on microdata (e.g. analysis in a user-defined area).

24. In Poland, geocoding process of statistical information was conducted for the first time during preparatory work for the 2010/2011 census round. Due to the lack of equal quality of reference materials, different steps concerning spatial accuracy geocoding objects were adopted. One of many stages of census preparatory work directly related to geocoding is that in the pre-census round enumerators had to verify whether buildings or address points existed, supplement the list with missing address points, verify the correctness of address points and determine their locations (x,y coordinates). The census enumerators were provided with mobile (hand held) terminals containing an application for the pre-census round.

25. After 2011 census round Polish statistics was in possession of actual digital statistical division boundaries as well as complete field verified database of address locations for the entire country. Since then these data have been updated on a regular basis. Since then statistical address points (address points for building and dwelling locations) are updated within statistics four times a year. Statistical division boundaries are updated annually. Above mentioned spatial data will be used in the upcoming 2021 census round based on the experience from previous census.

26. The scope of geospatial and statistical data integration is very complex. The challenge is to understand how to achieve this integration in the most effective and consistent way. Developing a coherent and systematic approach of linking statistical and geospatial data requires considerable commitment and time.

27. The best way to achieve the consistent integration is having a common method of enabling statistical and administrative data geospatially preferably in connection with Global Statistical Geospatial Framework1 (GSGF), which enables comparisons within and between countries. GSGF framework consists five principles that are considered essential to integrate statistical and geospatial information. Moreover GSGF is a high-level framework which facilitates consistent production and integration approaches for geo-statistical information. It is generic and permits application of the framework principles to the local circumstance of individual countries.

28. The creation of a geocoding infrastructure for statistics and its integration into the statistical production process does not demand a complete redesign of enterprise architectures and statistical production

[1] https://www.efgs.info/information-base/global-statistical-geospatial-framework/

processes. Small and stepwise improvements are possible. However, integration of statistical and geospatial information is a cornerstone in the modernisation of official statistics.

V. Big data and other alternative data.

29. 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: is it possible to use these type of data in census?

30. 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.

31. Due to the fact that Big Data is a relatively new phenomenon, many countries, not only Poland, have not yet developed clear guidelines for collection, processing and sharing such data.

32. Another problem arising both on the international and Polish level is the lack of experts - so-called data scientists.

33. One cannot also forget that, although Big Data carry a lot of potential, it is not without disadvantages. They are based on the fact that they do not always have the desired quality and even the speed with which they appear and are processed does not give grounds to take them into account without a thorough examination of their quality.

34. One of the most difficult obstacles to use Big Data for the census as early as 2021 is the lack of legislation that allows collecting, analyzing and storing Big Data. This is a problem affecting all countries, although to varying degrees. Therefore, those who cope better share their experiences with colleagues from other countries.

35. Lack of methodology, experts and relevant legal provisions are factors preventing the use of Big Data for the census as early as 2021, but this does not mean that Big Data will not be used in subsequent editions of the census. This is a new field, but very dynamically developing. The combined forces of statisticians around the world provide a kind of guarantee that today all problems constituting a barrier will be solved in the near future.

36. As part of activities after the 2021 census, it is planned to use alternative data sources, including non-administrative data, with large data volumes. Therefore, it will be necessary to use new processing methods. Currently, the possibilities of using BIG DATA tools in the statistical production process are being analyzed. Machine Learning components are tested as part of work on improving the quality of administrative data.

V. Summary

37. In the age of information society, it seems that it would be a mistake not to take advantage of the opportunity that modern technologies offer.

38. Current and future work on the improvement and use of registers for statistical purposes makes it possible to assume that subsequent censuses will be feasible using registers in full or with a small proportion of respondents. In order to reduce the costs of the census and the burden on the respondent when information is collected on many topics in a traditional census environment, the data collection should be based on the administrative registers as the main sources.

39. Poland participates in the work of international groups on the methodology and technology of future censuses, supporting directions of activities on the modification of the thematic scope and increasing the scope of data sources.

40. Active international activities to increase the scope of information sources for the census may have the effect of using alternative sources such as Big Data after the census in 2021.