



Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses**Twenty-first Meeting**

Geneva, 18–20 September 2019

Item 4 of the provisional agenda

Research on the use of administrative data for censuses**The use of administrative data to build a list of the population of people to the census****Note by Statistics Poland****Summary*

This paper presents the methodology for creating a list of the population covered by the census based on data from administrative sources. Furthermore, a new, innovative approach is presented which, in addition to determining the population covered by the census, will enable generation of a specific information range, taking into account any cross-sections and the lowest possible territorial division (including 1 km²). The approach takes into account the specificity of registry data, including their complexity and quality. The paper presents both methodological and technical issues related to the logic of the creation of the population list. The sources of data are discussed as well as the methods for their use, together with the creation of quality reports.

* Prepared by Pawel Murawski.



I. Introduction

1. The census is the only survey from which it is possible to obtain detailed information on the population, its territorial distribution, demographic and social structure and professional migrations (in particular of employees), households and families, as well as the quantitative and qualitative status of housing and building resources.
2. This information is collected at all levels of the territorial division of the country: nationwide, regional and local. Thanks to this approach, based on the information gathered in the census, it is possible to capture an image of the society, to create a kind of photograph of the current state and condition of the society. In connection with the implementation of activities aimed at increasing the use of data from administrative sources, the population list will be created using only data from administrative sources. This paper presents the method for creating the list, discusses the data sources, and shows both methodological and technical issues related to the logic of the creation the population list.

II. Building the population list

3. The purpose of building the population list is to create a broad and comprehensive description of the population for the 2021 census and annual censuses.
4. The construction of the population list was based on the integration of data from many heterogeneous sources to obtain a consistent image of the collected data.
5. At this stage of the work, it was assumed that the “connection key” for combining data sources would be the personal identification number (PESEL), because this is the primary identifier found in most administrative sources.
6. Table 1 presents the administrative sources that were selected for the creation of the population list.

Table 1

Description of the administrative data sources used to build the population lists

<i>No.</i>	<i>Name of data</i>	<i>Data provider/data source</i>	<i>Content</i>
1	PESEL	Ministry of Digital Affairs/Register of personal identity numbers	Persons registered with permanent or temporary residence in Poland
2	KEP	Ministry of Finance/National Register of Taxpayers	Natural persons (taxpayers) Natural persons conducting business activity (taxpayers)
3	ZUS	The Polish Social Insurance Institution/Central Register of Insured Persons The Polish Social Insurance Institution/Central Register of Contribution Payers	Natural persons (insured persons) Natural persons (contribution payers)
4	KRUS	Agricultural Social Insurance Fund Agricultural Social Insurance Fund	Farmers, their spouses and household members working with them (insured) Persons receiving agricultural pensions and disability benefits from KRUS
5	NFZ	National Health Fund	Natural persons (beneficiaries)

<i>No.</i>	<i>Name of data</i>	<i>Data provider/data source</i>	<i>Content</i>
6	ARIMR	Agency for Restructuring and Modernization of Agriculture	Natural persons (farmers)
7	PESEL_WY MELD	Ministry of Digital Affairs/Register of personal identity numbers	People who have been checked out of the register without a known destination, who had a permanent place of residence in Poland before the check-out

A. Creation of a data set of people with unique personal identification numbers (PESEL number)

7. The first step in building the population list people is to collect and merge unique PESEL numbers from the administrative registers given in Table 1. All personal identification numbers assigned to persons undergo a correctness check.

8. Validation of the PESEL number consists of checking the number of characters in the identifier, compliance of the check digit and the possibility of occurrence of the date of birth in the calendar that is encoded in the identification number. The PESEL number validation algorithm includes automatic checking of past dates as well as future dates, including leap years.

9. In the list of people holding a personal identification number, it is necessary to calculate the variables gender and age due to the need for further validation necessary for the construction of the list.

B. Verification of selected groups of people

10. The second step in building the list of people concerns the verification of the following groups of people.

1. Deceased people

11. Appearing on the list of deceased persons is verified on the basis of the current and previous years' information received. As a result of the work carried out, all deceased persons identified in the list are clearly identified.

2. People aged 90 and over

12. Based on the calculated WIEK variable, persons aged 90 and more are subjected to analysis (a dataset of records of persons for whom death is unlikely to be confirmed by a death certificate). Such persons are verified correctly if their personal identification number (PESEL) appears in the PESEL or WYMELD datasets and in at least one of the other datasets.

3. People who have left Poland

13. To identify persons who have left Poland, a collection obtained as part of ongoing research is used: permanent residents of other countries, registered in Poland for temporary residence, who have reported a departure from Poland or whose declared period of stay in Poland has expired.

4. Residents and recipients of a pension outside of Poland

14. Based on insurance codes with the Central Register of Insured Persons, it is possible to identify records of people residing and receiving benefits in a Member State of the European Union other than Poland or a Member State of the European Free Trade Association (EFTA). Such cases are considered in the context of staying or not staying in the territory of Poland.

5. People appearing only in one register

15. Persons for whom (part of) the personal identification number is found only in one register.

6. People with a residence address outside Poland

16. Based on the dataset of the National Register of Taxpayers, people with a foreign address are marked with the code for the country of residence.

17. The verification of individual groups of persons in the list of persons entails marking the examined cases (points 1–6 above), which enables the identification of records and improves the process of their thorough verification in the inventory, and which makes it possible to make necessary corrections in the process of calculating the population.

III. Appending personal data and addressing

18. The third step in creating a list of people involves the process of appending names and surnames to persons and determining their address. This is an important element of creating the entire list for the purposes of the census in the context of data personalization and census implementation through a census questionnaire (authentication). In addition, personal data can be one of the alternative connection keys.

A. Name and surname

19. The assignment of the surname and first name is preceded by a search for cases of different surnames and forenames for the same personal identification number, including checking for repeated values in individual administrative registers and differences between registers. Individual registers may contain such errors, mainly cases of errors in individual letters, which are easily verifiable by reference to another register.

B. Establishing address data

20. Establishing address information consists of appending addresses to individuals making up the list of persons, and actions to check the correctness of solutions and algorithms regarding the selection of the address of residence.

21. In determining the address data for persons on the list, four types of addresses are taken into account: permanent residence, temporary residence, registered address and correspondence address, which are collected in the registers indicated in Table 2.

Table 2

The order in which addresses are joined

<i>No.</i>	<i>Source</i>	<i>Type of address</i>
1	KEP	address of residence (registration address for the last stage)
2	ZUS	address of residence (registered address, correspondence for the last stage)
3	PESEL	time address
4	PESEL	permanent address
5	KRUS	address of residence (correspondence address for the last stage)
6	ARiMR	address of residence (correspondence address for the last stage)
7	NFZ	address of residence

22. The relevance of the data and the reference in terms of address data determine the hierarchy of address selection from a specific dataset. The level of consistency is taken into account for all addresses.

Table 3
Level of consistency of address data

No.	Level of consistency
1	consistency to street level
2	consistency to the place (city) level
3	consistency to the level of the commune
4	lack of consistency

23. The process of adding addresses according to the algorithm was divided into four stages:

(a) First, selecting the address consistency to the street level successively from the datasets of KEP, ZUS and PESEL;

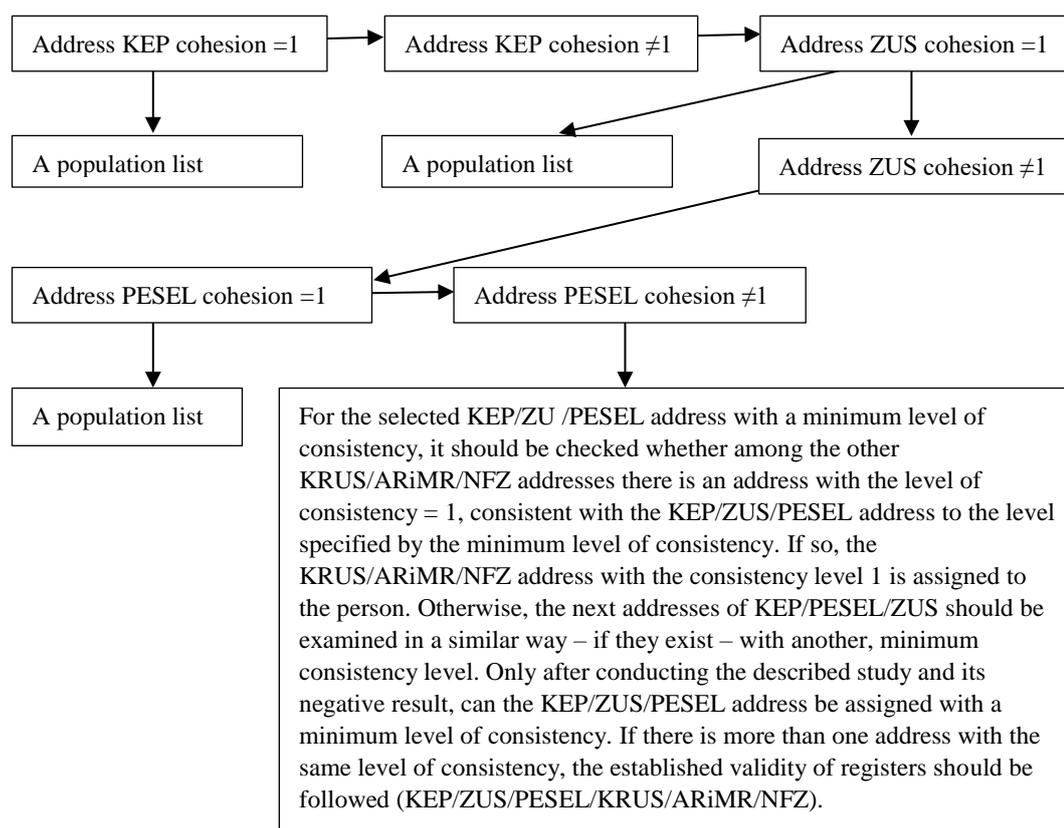
(b) Second, selecting addresses from the datasets of KEP, ZUS and PESEL with a minimum level of consistency (different from 1) and correction/supplementation based on the address from the datasets of KRUS, ARiMR and NFZ compliant with the address of KEP, ZUS and PESEL, to the level defined by the minimum level of consistency. Housing points for which a more complete residence record could not be found are included in the list of persons;

(c) Third, joining an address with a minimum level of consistency from the datasets of KRUS, ARiMR, and NFZ;

(d) Fourth, adding the so-called complementary addresses.

Graph 1

Simplified diagram of the address selection algorithm



IV. Conclusion

24. The presented method of building a list of people based solely on data from administrative sources will allow coverage of the full population of people who should be examined during the census. This allows full integration with other subjects, ensures high quality of data and is a fully automatic and repeatable process.
