United Nations ECE/CES/GE.41/2018/9

Economic and Social Council

Distr.: General
6 July 2018

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twentieth Meeting
Geneva, 26–28 September 2018
Item 2 of the provisional agenda
Methodology, new data sources including big data

Census Methodology in Estonia

Note by Statistics Estonia*

Summary

Statistics Estonia has developed an 'index methodology' to verify and specify the register data on the basis of a large number (24) of other registers and data sources. In the pilot census in 2019, this methodology will be tested in three particular cases – residency index, partnership index, and placement index.

Even though the general indexing principles have been established and model parameters have undergone empirical assessment, the methodology itself is still developing and new signs can be added depending on new information (incl. big data) becoming available.

* Prepared by Diana Beltadze and Ene-Margit Tiit.

1811154

I. Introduction

1. A register-based census in Estonia means that the dataset is extracted from registers. Before using registers' data as census data, the quality of data is verified in reference to basic statistical criteria. When data are included in different registers, they can be used for verifying the quality of data, on the one hand, and for selecting the most reliable values in accordance with the developed methodological rules, on the other hand.

2. Generally, census characteristics cannot be acquired directly from registers, because registers have been designed for other, non-statistical purposes and most of the definitions used are different from statistical definitions. It means that data from multiple registers have to be used in order to form certain census characteristics (e.g., the characteristic of 'activity status' requires data from more than 10 registers), while some characteristics are covered by duplicate information in several registers (Lehto, 2018).

3. The methodologists of census have solved the following main problems connected with the forming of census characteristics:

(a) Analysing the relationship of census definitions and value scales with the definitions and value scales used in registers;

(b) Testing the quality of registers and making efforts to urge register holders to eliminate any shortcomings;

(c) Determining the number of registers required for forming and, if necessary, verifying each census characteristic and, if some characteristics were not covered by any registers, taking steps to ensure creation of a respective register or register part;

(d) Establishing optimal rules for forming each census characteristic based on register data and creating the necessary software, using the quality of output characteristics as the basis of optimisation;

(e) Designing a strategy for using alternative rules when data gaps prevent the use of the optimal rule, as well as for imputation of values based on statistical or logical rules in some cases.

4. The biggest problem for a register-based census is the difference between registered and actual places of residence. This fact affects the breakdown of the lowest level of the place of usual residence (municipality) and all household and family characteristics.

5. There are few options to improve the situation.

6. The formation and structure of households could be solved by identifying quasi partners based on registers and other additional data sources e.g. big data (Tiit, 2017), similarly to the procedure of register-based determination of residency (Maasing, 2016).

7. How is it possible to correct registers using only register data? The answer comes from an old tradition of statistics – using repeated measurements allows to make the measurement results more precise. In a similar way, using a large number of registers, it is possible to improve the quality of administrative data.

II. Index-based methodology

A. Index-based methodology for register-based census

8. There are three different indexes worked out for a register-based census: residency index, partnership index and placement index (EU Grant 2016).

9. From each register containing information about people living in the country, it is possible to get signs which are useful for making decisions about persons. By the way, big

data can be adopted as additional alternative data sources that would help to improve the quality of census results or validate the obtained results.

10. In 2017, two pilot studies were conducted to test opportunities to specify the actual place of residence by using mobile positioning data and electricity consumption data to specify dwelling occupancy.

11. After the implementation of the partnership index, there will be a need to rearrange individuals' places of residence in the statistical register; electricity consumption data will be useful as well.

12. In connection to big data, the main focus of big data analysis is currently on organising data and preparing the data for analysis with classical methods.

13. New data categories and data formats require improvements in methodology and new methodological approaches. The data analysis methodology is significantly affected by calculation possibilities as well as opportunities to apply more and more complex and resource-demanding calculations.

B. Residency index

14. The residency index has been used in Statistics Estonia since 2015. The official population size and external migration size are calculated using the index. Also, the size of transnational population has been estimated using the residency index. In the residency index, about 20 registers and sub-registers are used; in time, the number of registers increases. The estimated inclusion and exclusion errors were smaller than 3% in 2013 and smaller than 1% in 2017; the new check is in process (see more on the website http://www.stat.ee/population-census).

C. Partnership index

15. The partnership index was calculated for the first time in 2017, using about 10 signs from 7 registers. Additionally, some continuous explanatory variables using time were created (such as the age of the youngest child, duration of marriage and the age difference of partners). The methodology used for determination of quasi-partnership is based on the dispersed set theory and it enables us to define a partnership index. The partnership index is used to isolate quasi-couples from the total population of couples. This is done by using partnership markers found in registers and other data sources (Tiit, 2017).

16. For procedures were used for estimating weights – logistic regression analysis, linear discriminant analysis, weights calculated by formula and their logarithms. In all the cases, the sum of inclusion and exclusion errors was 15% (in the optimal case, inclusion error was about *5%)(ibid).*

17. As the calculations were made for the first time, it was not possible to use the stability term of the formula. By using the partnership index and information from the Population Register, where all child-parent connections are fixed, it is possible to create all nuclear families. For forming households, some additional information is needed. The partnership index is established using partnership markers.

18. Partnership markers are binary attributes, which have the value of 1 if the respective situation exists (statement is valid) for the couple or the value of 0 if the respective situation does not exist (statement is invalid). The following partnership markers are used (Tiit, Visk, 2018):

(a) Have at least one child together;

(b) Have shared a place of residence;

(c) Are married or in a registered partnership;

(d) Have joint ownership of real property (dwelling);

(e) Have at least one shared (primary or secondary) dwelling (according to the criterion, they cannot have a shared main dwelling);

(f) Have filed a joint income tax return;

(g) Have shared parental benefits / parental leave;

(h) Have a shared or often-visited place of residence, which is not the main place of residence for either partner and which is identified based on indirect information, big data;

(i) Use (at least partially) a shared home postal address.

19. This is not an exhaustive list of potential partnership markers – the more markers we can include, the more accurate judgments (with lower probability of error) we can make.

D. Placement index

20. The placement index is a part of the general index methodology, and has not been used yet in Estonia. The common scheme of using indexes is the following:

(a) Check the residency of persons and in the future consider residents only;

(b) Form partnership couples and families.

21. For each family and single person, a placement index will be computed as a weighted sum of all the placement signs connecting a person (persons in the household) to particular living quarters. It is possible for some persons to have indexes connecting them to different living quarters, and particular living quarters can also be connected with several families. We suppose that also for those who do not have any signs of placement, a "place of existence" is fixed in registers, i.e. a city or municipality where he/she probably lives.

22. For all families, the suitability characteristic is calculated on the basis of family size and characteristics of family members (sex, age, status). For creating this characteristic, expert estimates and empirical data on satisfaction of families with their living conditions are used.

23. In a similar way, characteristics are created for all living quarters, containing information about their size, number of rooms, etc. The placement index connecting a family to a dwelling will be supplemented with a term characterising the suitability of the living quarters for the particular family. In decision-making, also this supplementary term will be used (EU grant 2016).

24. In general, each dwelling will be connected to one family, but there are some exceptions:

(a) In one dwelling there may live two families, if they are relatives (two generations);

(b) It is possible that in the living quarters of a family there may live some single persons (e.g. students or other subtenants).

25. People/families who do not have any placement signs will be connected to empty dwellings that are suitable for them and are situated near their place of existence. If there

are several possibilities, random choice will be used to assign a dwelling for each person/family.

26. While working out the methodology for the register – based on a census it became obvious that the list of data sources is subject to change. The lists of signs of partnership and signs of life are never final. We keep looking for new data sources and we have to be ready when some sources are discontinued.

E. Testing indexes

27. The residency index was tested during the first pilot census in 2016.

28. The aim of the first pilot census of the register-based population and housing census was to test the quality of the registers used to form the census characteristics, the functioning of the methodology, and readiness of the support software for a register-based census.

29. Like a regular census, a pilot census was conducted by establishing a moment of census (31.12.2015) and forming a census database as of that moment.

30. The created data tables were verified and the quality level of data was determined in relation to the established criteria and the quality requirements for statistics.

31. Different weaknesses detected in the trial census will be addressed with relevant measures and the results will be tested in a second pilot census in 2019.

32. The results of the first pilot census confirm that using the residency index is suitable for a register-based census. The pilot census covered the census characteristics, which are mandatory for the Member States of the European Union. In addition, the analysis included the characteristics of ‘ethnicity’ and ‘native language’.

33. The partnership index and placement index will be tested twice. The quality of these indexes will be tested during a sample survey (2% from total population) in 2018 and the next test will be during the second pilot census in 2019.

III. Using mobile data as an additional source for census

A. Using mobile data as additional sources for census

34. Statistics Estonia has started a pilot project for testing the possibility to use mobile positioning data (MPD) for a register based census. The problem we wanted to solve was the difference between registered and actual places of residence, which cause biases in population statistics.

35. An aim of the pilot study was to test the possibilities to validate existing addresses. The pilot study took place from April 2017 to May 2018. Results were presented in May 2018.

36. Steps of the pilot study and timetable:

(a) Volunteers were asked to participate in the pilot study during spring 2017;

(b) A set of potential addresses was created for each participant based on registers during summer 2017;

(c) Obtaining from mobile network operators (Telia, Elisa, Tele 2) individually identifiable data on mobile operations performed on the phones of consenting persons from August 2017 until March 2018;

(d) Data processing (conversion to a format required for calculation of anchor points) was performed in early spring 2018;

(e) The resulting data were used for calculating relevant locations, i.e., home anchor point and home-workplace, working time, secondary home and other regular anchor points in April 2018;

(f) Analysis of the accuracy of identified home anchor points and home-workplace anchor points in comparison to actual home address data provided by the persons was carried out in April;

(g) Analysis of the accuracy indicators of results and sources of errors was performed in April 2018;

(h) Methods were provided for the identification and validation of the actual residence from among the residence addresses in the Population Register and other addresses found in registers in May 2018.

37. The analysis entailed calculation of home anchor points by applying Positium's standard algorithms on passive MPD, detection of anchor point overlaps with actual residence addresses provided by the respondents, and comparison of the parameters of mobile data with address matches. The percentage of correctly identified anchor points was calculated for administrative units. Definitions were based on the range areas of anchor points and their centroid locations (Positium, 2018).

38. The comparative analysis by the Positium provided process comprised the following:

(a) Pseudonymisation of the data obtained from operators, separation of individually identifiable attributes (name, phone number) from the processed dataset;

(b) Verification and preparation of raw data;

(c) Geocoding of residence addresses;

(d) Calculation of anchor points for involved persons;

(e) Comparison of anchor points and geocoded home addresses;

(f) Statistical analysis of the results: comparisons between different attributes of mobile operations and the percentage of correctly identified anchor points;

(g) Comparison of MPD and home addresses.

39. The locations of home anchor points calculated on the basis of MPD were compared to the home addresses specified on the consent forms. The overlap between the range areas of home anchor points or home-workplace anchor points and the location of the range area centroids in the same administrative unit as residential address produced the following results:

(a) Actual place of residence is inside the home or work-home anchor point polygon – 82%;

(b) Actual place of residence and anchor point:

(i) In the same county – 97%;

(ii) In the same municipality – 87%;

(iii) In the same settlement – 67%.

40. For some individuals, anchor points could not be calculated using the MPD method and, consequently, MPD could not be used for their residence validation due to their limited telephone use or technical issues associated with processing and interpretation of data, including calculation of anchor points.

41. The disadvantage of observing settlement units in relation to the centroid of an anchor point range area stems from the fact that, normally, the range area of a mobile mast includes several settlement units (especially in sparsely populated areas) and this does not necessarily mean that the anchor point has been incorrectly calculated (Positium 2018).

42. The next step was validation of PR address using two methods of mobile data: distance-based validation (based on distance of MBD range centroids) and range-based validation (based on the positioning of a register address in the range area of an anchor point, with additional attributes).

43. The goal of the methodology was to identify and validate the 'actual' residence address from the residence addresses in the Population Register and any other addresses found in registers.

44. Distance-based validation of residence would be the simplest option, but this method does not facilitate estimation of the reliability (level) of validation *(ibid)*.

45. The distance-based validation method does not take into account mobile network range indicators.

46. For this reason, distance-based validation is usable with simple scenarios but cannot provide reliable answers in more complex situations.

47. Range-based validation enables to take into account the position of a register address in relation to the range area. This facilitates estimation of the reliability of each validation and can provide lower-level validation even for more complex and boundary cases.

48. It can be pointed out that the anchor point identification methodology contains some potential for errors, especially in relation to individuals with irregular behaviour patterns (Sõstra, 2018).

49. As a possible data source we considered using big data to determine partnership, but this is not crucial and the use of such data would be possible only after serious work evaluating data quality, which has not been done so far.

IV. Concept of the Second Pilot Census in 2019

50. As it was mentioned, Estonia's greatest problem is the inaccuracy of residence data in the Population Register. Therefore, it is necessary to develop an index-based methodology to verify the register data on the basis of a large number of other registers and other data sources.

51. During the second pilot census, this methodology will be tested in three particular cases – residency index, partnership index, and placement index.

52. The main concept of the second pilot census is based on two possible scenarios for the next population and housing census, with relevant empirical data collected for both options.

A. Option A

53. Piloting of a full-scale register-based census where all EU mandatory output characteristics are calculated on the basis of register information.

54. This option facilitates testing of (EU Grant 2016):

(a) Availability of information in registers and transportability of the data;

(b) Quality and coverage of the register information in relation to the total population;

(c) Performance and accuracy of the algorithms developed for the calculation of census characteristics;

(d) Capacity of model-based indexes (residency index, partnership and location index) to generate estimates that reflect the actual situation.

55. The outputs of the pilot census include the most relevant hypercubes, which will be compared to the respective cubes determined on the basis of data from PHC 2011. This helps to highlight any developments in the recent period and to test the adequacy of detailed information obtained from register data.

56. The quality of the results of the pilot census will be assessed according to developed rules and norms both with regard to individual characteristics as well as sets of characteristics (cubes and marginal cubes) *(ibid)*.

57. Findings of the pilot census:

(a) If the pilot census produces adequate results and outputs that meet the international requirements, it means that a register-based census is feasible in Estonia;

(b) If the results indicate that some census characteristics:

(i) Cannot be calculated on the basis of registers; or

(ii) Coverage or quality does not meet international requirements,

then Option B will be implemented.

B. Option B

58. Option B will be implemented if registers cannot guarantee the required level of accuracy for some census characteristics (international practice indicates that this number can be between 3 and 5). The methodological work performed on census so far indicates that the number of non-conforming characteristics will certainly not be higher than that. For the remaining characteristics, the register-based census methodology described under Option A will be applied. The characteristics that cannot be estimated on the basis of registers with sufficient accuracy will be handled separately *(ibid)*.

59. For such characteristics, the data collected with large-scale surveys of Statistics Estonia in the past year(s) will be used to impute the values of missing characteristics according to a suitable statistical imputation (prediction) procedure. Imputation ensures consistency between all hypercubes and marginal cubes unlike, where necessary information cannot be obtained on small groups of people.

60. This combined methodology conforms to international requirements. It also retains the current person-based approach to population statistics.

61. The main body of work in the pilot census will be based on Option A. In addition, the pilot census will be used to identify suitable imputation methods and to test imputation performance with at least one characteristic. The result will be compared with the register-based data distribution for the same characteristic *(ibid).*

VII. Conclusion

62. The greatest problem during the census preparatory works is the inaccuracy of residence data in the Population Register. This has forced Statistics Estonia to develop an 'index based methodology' to verify and specify the register data on the basis of a large number of other registers (24) and other data sources (big data).

63. During the pilot census in 2019, the new methodology for register-based census in Estonia will be tested in three particular cases – residency index, partnership index, and placement index. All three indexes use Estonia's administrative databases as sources of information, which can be combined to form an interoperative data system with common identifiers.

64. Assuming that, in the present day, a person living in Estonia inevitably leaves certain traces of activity in the form of records in different databases, it is possible to verify the person's residence in the country, as well as connections between persons and their locations, on an annual basis. Such verification is based on signs of life, signs of partnership and signs of placement that are recorded in registers every year.

65. The annual indexes are established as linear combinations of the respective signs, which makes it possible to trace the change in a person's status in different years. The indexes are calculated for all persons.

66. The methodology itself is still developing and new signs can be added depending on new information (incl. big data) becoming available. The accuracy of the index-based estimates is assessed through use and additional surveys, and the results are provided with potential estimation error values. Addition of new information (further signs) will result in consistent improvement of the accuracy of index-based estimates.

References

Tiit, E.-M. (2015). Residence testing using registers – conceptual and methodological problems. Presentation at 4th Baltic-Nordic Conference on Survey Statistics. [www] https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/17062664 Tiit_Abstract.pdf.

Tiit, E.-M., Maasing, E. (2016). Residency index and its applications in censuses and population statistics. Quarterly Bulletin of Statistics Estonia 3/16, pp. 53–60.

Tiit, E.-M., Vähi, M. (2017). Indexes in demographic statistics: a methodology using nonstandard information for solving critical problems. Papers on Anthropology XXVI/1, pp. 72–87.

Tiit, E.-M., Visk, H., Levenko, V. (2018). Partnership index. Eesti Statistika Kvartalikiri 2/18.

Eurostat Grant 2016. (2018). Improvement of the use of administrative sources (ESS.VIP ADMIN WP6 pilot studies and applications). Methodology report.

Sõstra, K. (2018). Presentation about results of the pilot study of mobile positioning data.

Positium (2018). Report about of the pilot study of mobile positioning data.

Estonia's first register-based trial census. 2016 [http://www.stat.ee/population-census]