Statistical Journal of the IAOS 36 (2020) 159–164
DOI 10.3233/SJI-190585
IOS Press

Developing methodology for the register-based census in Estonia

Diana Beltadze
Statistics Estonia, 51 Tatari Street, Tallinn, 10134, Estonia
E-mail: diana.beltadze@stat.ee

Abstract. The next census in Estonia is planned to be register-based – all the necessary information would be derived from 25 registers. During the preparation period, the algorithms for creating census variables from register variables were prepared and tested and criteria for quality checks were developed. The first step for the state was to create a legal basis for using administrative data, with the aim to produce statistics. It was also necessary to link all administrative registers with each other by identifiers (ID codes of persons, enterprises and addresses) and form a uniform system. The existence of a rapid and secure data exchange and transportation environment was also required. Big efforts were made to achieve good mutual understanding and collaboration between Statistics Estonia and register holders.

Keywords: Census, registers system, cross validation of registers, indexes

1. Introduction

Change in the census method in countries has been quite slow. Since the 1970 round of censuses, statisticians in the whole world have considered it important to search for alternative methods to traditional census, in order to reduce the cost of censuses and the burden on respondents, but maintaining the necessary quality of results. However, the introduction of new methods has proved difficult. The most innovative have been Nordic countries. There, the tradition of population registers, developed from the lists of congregation members, was already several hundred years old. The first non-traditional censuses were organised in Denmark in 1981, with all the information extracted from the population and other registers.

When Estonia was occupied by the Soviet Union, all statistical activities, including censuses, were carried out by central command and in a unified form, prescribed by Moscow, and no innovation was possible. After Estonia regained independence in 1991, new thinking started to emerge also in statistics. All Estonian residents were given unique ID codes and the first administrative registers were established. Relying on preliminary lists of population counts, the Estonian population register was created in 2002. The idea of a register-based census in Estonia was born at that moment.

The potential of using registers for census was extensively studied by Statistics Estonia. Although there were some other databases and registers in the country (for buildings, for instance), their quality was not checked. The data inclusion rate of these databases was too low, the concepts did not correspond to the needs of the population census, and were not compatible. Many of the necessary characteristics were unavailable in the registers. In addition, the legislative base was often inadequate, and there was opposition in the public – having experienced persecution by the Soviet state, there was fear of a new "super-register". The improvement of the registers on the basis of knowledge gained in the latest census became an important task, but the rule of unidirectional moving of microdata collected by statistical offices was kept, which means that the census data were not rendered to registers.

Before selecting the method for the 2011 census in Estonia, a decision had to be made regarding the long-term development of censuses in Estonia, i.e. whether to move as fast as possible towards a fully register-based census, or develop the traditional census through the use of new technologies and other innovations.

Statistics Estonia decided to increase the use of administrative registers for statistical purposes in different fields. The target was to prepare for a register-based census using information technology solutions to maximise automated data collection.

For a register-based census in Estonia, a number of conditions had to be met at national level, the most important of which are:

1) a legislative basis supporting the creation and continuous updating of the registers, providing access to registers for the production of statistics and the right to combine data from different data sources;
2) availability of a population register of sufficient quality;
3) availability of building and dwelling registers of sufficient quality;
4) availability of sufficiently high-quality registers or databases that allow determining economic activity of the population (tax, pension, social protection and employment service registers and a business register);
5) a link between various registers and databases, which requires the adoption of identifiers (unique identifiers for persons, dwellings and buildings, and also for enterprises);
6) adequate harmonisation of register and census concepts, i.e. definitions in registers are useable for defining census concepts;
7) technical environment for rapid and secure data transportation from registers (via X-Road, which is the data exchange layer for information systems, a technological and organizational environment enabling secure Internet-based data exchange between information systems);
8) good cooperation between the producers of statistics, register holders and administrative bodies.

It is clear that registers can be of high quality when people accept the need for their data to be held in registers and understand that this is useful and necessary for them. However, a trusting relationship between the citizens and the state is not as strong in transition countries (Central and Eastern European countries) as, for instance, in the Nordic countries [1]. As a result, many people do not want to give adequate data (real addresses, for instance) to registers. Statisticians have to make a big effort to find defective data among the registered data and improve them for statistical use. The management of high-quality registers also requires sufficient financial resources for their continuous development and updating.

2. Preparation for the register-based census in Estonia in 1996–2008

The possibility of a register-based census in Estonia has been considered since 2002. Before that, in late 1990s, it was rather non-realistic. In general terms, it can be said that the problem is as old as the national registers of Estonia. The issue that has been discussed more is the use of standards (address data, classifications). It has been stressed that selecting the register-based census methodology requires that all mandatory census variables are to be covered by registers, that there is a system whereby all objects observed by census have been identified and that address data are used.

The issue of registers emerged in Statistics Estonia's correspondence with authorities again in 2005, when the quality of registers was assessed. One of the reasons why the issue was not addressed earlier was the strict legal environment for producing statistics, e.g. it was prohibited by law to use the microdata of the 2000 census for statistical analysis (this applied until the entry into force of the new act in 2010). This meant that data validation work using register data was not possible in Statistics Estonia.

A new initiative was launched in 2007. This involved preparations for a new census round. Statistics Estonia analysed the availability of the population and housing census indicators in state databases (11) on the basis of a self-assessment questionnaire of the registers [2]. The study concluded that the use of register data, considering the compiled census programme, was not feasible due to the fact that databases and registers did not have sufficient information for census variables (*ibid*). The main shortcomings concerned the mandatory output of the European Union (EC, 2008) as well as additional variables needed by Estonian users. There was a complete lack of data on household composition and actual family status, but also on working time, religion, foreign language skills and agricultural activities of households required by different groups of Estonian users. Data on occupation, mother tongue, links between households and household members, the number of children born to a woman, educational attainment, migration and the place of birth of parents were under-covered. The data indicating living conditions was lacking. Metadata in registers were also incomplete. It was unknown whether the actual place of residence and the registered place of residence of persons coincided. There was no overview of people living in institutions (monasteries, children's homes, etc.). It was not possible to link the population regis-

Table 1
Person, household and dwelling characteristics in registers in 2013

Assessment of coverage in registers	Mandatory census variables
A – complete	Sex, age, country of birth, nationality.
B – partial	Labour status, occupational status.
	Useful area/number of rooms, building type of dwelling.
C – limited	Permanent residence, residence abroad and year of arrival, previous residence, relationships between household members.
	Housing arrangements, type of dwelling, basis for the use of dwelling, ownership status, number of persons living in the dwelling, technical characteristics of the dwelling (water supply system, toilet, washing facilities, heating type, time of construction).
D – none	Occupation, place of work.

Source: Project "Register-based census methodology report" [1].

ter to the register of construction works and buildings; hence it was possible that a person was registered in a dwelling that did not exist.

It was difficult to explain to the public and authorities that data needed from the census were not available or did not have the required quality in the registers and that combining the registers with survey data does not provide a satisfactory result. After a while, though, it became clear that it was too early to organise a register-based census in Estonia in 2011.

In conclusion for the period 1996–2008, it should be noted that there were two approaches to the use of registers for census: a pessimistic one, emphasising a lack of data quality and big data gaps in the registers, and an optimistic one, which was oriented towards the use of registers. The optimistic approach highlighted the possibility to save on population and census takers burden and also census costs by using modern information technology tools and possibilities to integrate data of various registers, reorganising the dispersed way in which information was handled by different authorities [5]. The view of the representatives of the pessimistic approach sprang from the responsibility for the quality of census results, gaps in legislation, lacking opportunities to update hardware and software platforms, including databases and systems, for the production of statistics. At the same time, there was a desire and willingness to cooperate more closely with various authorities and agencies to study the data of registers and find ways for improving the situation. In order to improve the data quality of registers, the producer of statistics recommended specifying data quality requirements and introducing auditing of the registers. It was also reminded that there was a need to ensure regular updating of the data in the registers and that it was crucial to harmonise the definitions and classifications used in different registers.

3. Preparation for the register-based census in 2009–2015

New impetus was given to the register-based census in the context of entry into force of the Official Statistics Act in 2010. Statistics Estonia was given the task to start preparations for a register-based census. The task of comparing 2011 census data with data in registers was stipulated by law (Official Statistics Act, chapter Census). The coverage of databases and registers was insufficient with regard to required indicators before the previous census [7]. It was settled that the assessment and improvement of the quality of registers using census data was a prerequisite for the transition to a fully register-based census. A comparative analysis of datasets was done in 2008–2013 after all census results were published. It was concluded based on the analysis that it was possible to get the needed information on most mandatory census variables by combining registers and 2011 census data. This concludes that registers could be used for a census. At the same time, similarly to the results of the methodological report in 2013 (see Table 1), the analysis results showed that in many cases the data quality might be insufficient.

In general, the inter-institutional methodology project resulted in a relatively negative assessment of Estonia's system of registers and raised questions about Estonia's ability to conduct a register-based census. Nonetheless, preparations for the register-based census in Statistics Estonia continued. The 2013 Register-based census methodology report [1] also refers to the need to start work to improve data quality in the registers; however, the question about how the collected data should be improved in the registers remained. Registers are document-based in Estonia, which sets a limit on their corrections, as there is a need for a legal basis. Legislation for this is complicated and the situation cannot be solved either by

the register keeper or Statistics Estonia. There were no good solutions to the issue at this point in time.

By 2014, as a result of intensive collaboration of statisticians and register holders, a new level had been reached in terms of the interoperability of state information systems:

1) almost all census characteristics, except occupation and place of work, were covered by registers;
2) there was a system of personal identification codes, on the basis of which all personal data registers had identifiers (ID codes) and the linking of personal data was possible;
3) there were codes for address objects and the inhabitants were linked to their dwellings by address codes and ID codes of persons. People were linked to the organisations where they were working using commercial register codes. Negotiations for amendments to the legislation were initiated to begin collecting data into the employment register (place of work and occupation) and the population register (residential address).

Considering the new situation, Statistics Estonia decided to start preparing for a register-based census and test the situation in two trial censuses in 2016 and 2019.

New goals were set to:

- resolve questions of the substantive/legal/technical possibility of a register-based census;
- settle the relationship between the definitions of registers and census variables;
- create a concept and plan for capturing census data;
- develop a methodology for determining the population of the census, i.e. who should be enumerated;
- develop a framework for quality improvement that would ensure coverage, accuracy, regular updating of data in registers.

In conclusion, during the period of discussions about the register-based census in 2009–2015, progress was made towards networking and interoperability of information systems. An address standard was adopted by the population register; data quality was measured both by database keepers and Statistics Estonia; there was a secure data exchange environment; registers had owners and the owners had tasks; census data could be used for data methodology work and quality standards were established for data.

4. Preparations in 2015–2019

The methodology report and results of the last census were analysed and a decision was made in Statistics Estonia to continue developing the register-based methodology within the EU grant works in spite of the pessimistic prospect for the register-based census from the point of view of the authors of the report [1].

From 2015 to 2019, the methodology team for the register-based census in Statistics Estonia worked in the following main directions:

1. cooperation with state registers to verify the quality of the data therein, identify shortcomings and support efforts to improve the quality;
2. development and testing of algorithms for the calculation of census variables based on the information in registers;
3. development of the methodology (indexes) for correcting inaccuracies in register data using models based on the information obtained from multiple registers and other sources of information.

The main concern in a register-based census pertains to incorrect data submitted by the population. Estonia's greatest problem in this respect is the inaccuracy of residence data in the population register.

This has forced Statistics Estonia to develop an "index methodology" to verify and specify the register data on the basis of a large number of other registers and data sources. 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. Index is defined yearly as a linear combination of signs received during the year. The value of an index varies between 0 and 1, and it can be considered a probability of a positive event: the person is a resident, the pair of persons constitute a partnership couple, the household lives in the given dwelling. To make a decision, a threshold is used: if the index value is higher than the threshold, the event is considered positive. In defining the index, there are two crucial questions:

1) Defining the weights of signs. In principle, the task can be solved by using some multivariate procedure (logistic regression, discrimination) or machine learning procedure, if there exists a good and reliable training sample.
2) Defining the threshold that allows to make a decision. To calculate a threshold, empirical training data are used, and the value of the threshold

is calculated in a way that the inclusion error and exclusion error are in balance and as minimal as possible [11].

These indexes use Estonia's system of administrative registers, which uses common identifiers and allows linking and combining the data of different registers. 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 [12]. The annual indexes are established as linear combinations of the respective signs, which makes it possible to trace yearly the change in a person's residence or partnership status and dwelling address.

The indexes are calculated for all persons who have ever received an Estonian personal identification code. This makes it possible to monitor transnational persons who have left Estonia, including to detect whether they have returned, or how trans-boundary commuters move between their homeland and other countries [13].

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 (including big data) becoming available [14]. The accuracy of index-based estimates has been assessed through the use of additional surveys, and the first results are very good in the case of the residency index, good in the case of the partnership index and satisfactory in the case of the placement index. Adding new information (further signs) will result in the consistent improvement of the accuracy of index-based estimates.

The index-based methodology has been presented in research articles and at international conferences. Population statisticians of several countries who face similar problems have expressed interest in the practical applicability of the indexes. The baseline situation for the register-based census can be quite diverse in countries but there are international requirements and standards for the outputs of register-based population and housing censuses. These requirements are the same, irrespective of the particular census methodology. Considering this background, it is very important to plan and execute the necessary number of successful pilot censuses before the first register-based population and housing census.

The rehearsal in 2019 was to pilot a full-scale register-based census where nearly all output variables are calculated on the basis of register information.

This option facilitates the testing of:

- the availability of information in registers and transportability of the data;
- the quality and coverage of the register information in relation to the total population;
- the performance and accuracy of the algorithms developed for the calculation of census characteristics;
- the capacity of model-based indexes (residency index, partnership index) to generate estimates that reflect the actual situation.

The quality of the results of the pilot census were assessed according to the developed rules and norms, with regard to all mandatory output variables as well as sets of variables (cubes and marginal cubes). The results of the second pilot census were reassuring.

At the end of October in 2019, the census variables were agreed upon in the census committee of the Government of Estonia. The variables included those based on the needs of the users of statistics in Estonia, including researchers. Not all these new variables can be compiled on the basis of registers. Therefore, it was recommended that Statistics Estonia review its decision about register-based census approach for the next census. Another option for the census will be implemented.

Estonian users have been very demanding and critical towards the quality of census results. There are also research groups taking interest in different facets of society and they wish to obtain information on health, religion, command of languages, ancestry, etc. from the census. A large part of this information is also available from registers, but the extraction needs additional efforts, e.g. a data-mining methodology.

Indexes were also tested using a special survey in 2018, which demonstrated that despite the fact that indexes improve deficiencies in the data, not all of them are 100% reliable. Partnership index should be improved in the case of young couples. Also determining the real place of residence of young people (students) was determined to be rather difficult.

The strategic risk has been emphasised that the results of the register-based census will not meet the needs of Estonian users, because it is difficult to find a workable solution to improve the data quality of the place of residence in the register.

However, there is a possibility to use big data. For instance, it is possible to check if a dwelling is inhab-

ited by using data on electricity use – preliminary research in this direction has been made in Statistics Estonia, but these data have not yet been added to the set of signs of placement for matching people with their actual address. Another possibility is to use mobile phone data. In Estonia, these data have been used to analyse migration flows. Although the data are not identifiable, they can be used for checking the number of inhabitants of dwellings and farms.

5. Conclusion

Developing methodology for the register-based census is time-consuming. Before using data from registers for censuses, the quality of the data has to be verified in reference to basic statistical criteria. When new data are included in different registers and datasets, they can be used for verifying the quality of existing data, on the one hand, and for selecting the most reliable values in accordance with the developed methodological rules, on the other hand.

Generally, census variables cannot be acquired directly from registers, because registers have been designed for other, non-statistical purposes and most of the definitions used differ from statistical definitions. This means that data from multiple registers have to be used in order to form certain census variables (e.g. the variable of "activity status" requires data from more than 10 registers).

Methodologists have solved the main problems connected with the forming of census variables, correcting incorrect values of some variables. This work has demonstrated that a register-based population and housing census is feasible and the preparations for the census have been purposeful. But we cannot say that it is possible to conduct a solely register-based census in 2021. The biggest problem for Estonia is the difference between the registered and actual places of residence. This affects the breakdown at the lowest level of the place of usual residence (municipality) and all household and family characteristics. Along with data, also the methodology of census statistics is continuously developed, i.e. new possibilities will emerge for processing data. New data categories and data formats require improvements in the methodology and new methodological approaches.

References

[1] Puur, A., Sakkeus, L., Aben, S. (2013). Register-based census methodology report. In Estonian. Tallinn.
[2] Paut, H. (2007). Summary of the analysis of 11 registers. In Estonian. Tallinn.
[3] Paut, H. (2007). Summary of the analysis of 11 registers. In Estonian. Tallinn.
[4] Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses. Available from: https://eur-lex.europa.eu/legal-content/ET/TXT/?qid=1546954021572&uri=CELEX:32008R0763.
[5] Beltadze, D. (2016a). Securing interoperability for the first register-based census. Tallinn University.
[6] Official Statistics Act of Estonia. 2018. Chapter 4. Census. Available from https://www.riigiteataja.ee/en/eli/ee/506012015002/consolide/current.
[7] Beltadze, D. (2016b). Information technology and its impact on productivity. Paper for the workshop on the census, Geneva, Switzerland.
[8] Puur, A., Sakkeus, L., Aben, S. (2013). Register-based census methodology report. In Estonian. Tallinn.
[9] Puur, A., Sakkeus, L., Aben, S. (2013). Register-based census methodology report. In Estonian. Tallinn.
[10] Puur, A., Sakkeus, L., Aben, S. (2013). Register-based census methodology report. In Estonian. Tallinn.
[11] Tiit, E.-M., Vähi, M. (2017). Indexes in demographic statistics: A methodology using nonstandard information for solving critical problems. Papers on Anthropology XXVI/1, 72–87.
[12] Tiit, E.-M., Maasing, E. (2016b). Residency index and its applications in censuses and population statistics. Quarterly Bulletin of Statistics Estonia, 3, 53–60.
[13] Tiit, E.-M., Visk, H., Levenko, V. (2018). Partnership index. Bulletin of Statistics Estonia No 2.
[14] Beltadze, D. (2018). Census Methodology in Estonia. Paper for the workshop on the census, Geneva, Switzerland.