Seminar on Registers in Statistics - methodology and quality
21 - 23 May, 2007 Helsinki

Usage of registers' data in censuses

Ene-Margit Tiit
Statistics Estonia; University of Tartu
ene.tiit@ut.ee; ene.tiit@stst.ee

Chapter 1. The aims and resources of two statistical collections – census and registers

The Censuses are the largest and most important statistical collections undertaken by any country. The census must measure all important dimensions of the society; it is like a snapshot of the whole society with all its members and smallest details. The history of censuses started more than 5 thousand years ago and, in general, the methodology of censuses is rather conservative to guarantee the comparability of data across long-lasting time series and different countries [1, 2]. But – there are objective factors and reasons demanding changes in the ideology and practical organization of censuses in the 21st century.

In the modern society it became evident that a series of periodically made snapshots does not give enough information to follow the development of the society and to make adequate decisions. The permanent need for information was the reason why during the last half-century, especially during the last decades many different registers have been created in most developed countries. Many of them cover the topics measured by censuses and, indeed, some of them have also the pre-history of several hundred years. In fact, the registers form a modern collection of statistical data that is continuously updated. As the register data reflects (at least principally) the situation of the society at the given moment, it forms a tool for monitoring the development of the society.

The problem we will regard here is – how much the registers can be used to solve the problems of censuses and can they improve the quality of census data? Can the registers substitute the traditional censuses and which are the preconditions of that?

The recent changes in the society that influence the data collection

During the last several decades the society in the whole world has changed dramatically. That means, the object of census – the population – has been changed, whereas the change has not been linear increase or decrease, as it was earlier. Several new dimensions have arisen and they need new, more flexible measurement instruments, compared with traditional censuses. Especially in Europe – both in "old, i.e. western" and "new", i.e. eastern Europe the changes have been remarkable. The leading terms characterizing the society today are ***mobility*** and ***diversity***.

- Mobility in the sense of living place means that many households have several living places (in the same or different settlement or country which they use either simultaneously or seasonally;
- Also it is quite common that a people works in one city, studies in another city, has home somewhere in countryside and spends vacation somewhere abroad.
- Mobility in the sense of family or household relations means that its is very difficult to define which couple is a family/ household and which is only a short-time union (often the partners themselves have different understanding in the sense);
- There are more and more children whose parents live in different places and children circulate between both parents, sometimes also between parents and grandparents.
- But there are also some features of "reverse mobility" – there are more and more people working at home and communicating with employer via internet, whereby the last one might be situated in another city / country.

Diversity means that it is difficult to define a finite set of fixed standard “shells” to cover all people of the society/ in a country.

- There are very different family forms: homo- and hetero couples, married and unmarried, couples living together and living in different addresses, having common, separate and adopted children;
- The societies are more and more multinational, multiethnic, and multicultural and in spite of all possible education programs, there are quite a lot of people who cannot and do not like to communicate in official language of the country.
- But, besides young people using all modern possibilities of communication there still exist elderly people, who like their traditional living style in countryside without computers, internet and mobile phones. The share of elderly people is increasing and that means also that the diversity in age is increasing.

Most of these problems are not new in principle, but since they were rather exceptions that formed in all statistical tables a shell named “others” and share less than a part of percent, they were not used in common statistics. Nowadays all these features are increasing and cannot be ignored. That means they must be measured to have the possibility to forecast the rising trends in society.

New problems in organizing censuses

Nowadays the organization of a traditional census will face with long list of different conceptual, methodological and also economic problems. Most of them are new or are much more serious today than they were earlier.

The high expectations:

- The expectations of the quality of the results of census are very high: the results must be unbiased, with high coverage and response rate, punctual, correct and adequate.
- The number of dimensions to be measured increases steadily.
- Not only prevailing current status, but also growing tendencies in the society must be measured and documented.

The unwillingness of collaboration of population

- The people are less and less willing to answer to any questions concerning their private life.
- As time (also every minute) is nowadays very highly rated, many people do not want to communicate with the interviewer. They find senseless to say the things that they have already said (in former census). The longer the interview is, the larger is the probability of discontinuing it.

Conceptual problems – definitions of basic concepts

- As result of changes in society such concepts as family and household, (usual) living place, member of household, institutional household etc, must be rethought to cover better the new features and situations in the society. E.g., the common definitions of household (the set of people having common roof and budget) and usual living-place (the common living-place of a household) form a logical circle and exclude all households having several dwellings or members living apart.

Technical problems

- Because of high mobility there are big problems with reaching the respondents.
- The diversity of society members (including the homeless and other marginal groups) complicates communication of interviewer with respondents.

That means the organization of censuses must develop with the aim of meeting new demands and avoiding the new problems. One of the important approaches is ***the information economization principle.***

The information economization principle (IEP) in statistical data collection

The idea of the principle includes the following three steps:

- maximal use of existing information (registers, data-sets of recent census) in censuses;
- maximal check of existing/used information during the census;
- maximal use of information received in census to improve existing data-bases and registers.

The application of this very simple and seemingly rational principle has a lot of serious hindrances and problems. The most important problems are the following:

The juridical problems

- concerning the personal data protection rules;
- concerning the aims of census
- concerning the administration of registers and data-sets.

The conceptual problems

- Existence and adequacy of identifications in different data-sets:
  - ID-code of persons;
  - Address of dwellings;
- Sometimes definitions do not coincide in different registers/ registers and census instruments;
- Classifications might be different in different registers and census instruments.

The coverage problems in the sense of dimensions

- Some important dimensions to be measured via census are not covered by any register.

The coverage problems in the sense of population

- Some population groups belonging to the scope of census might not be accounted in registers (and vice versa).

The problems of quality of register data

- It may happen that the registers do not check/ monitor the quality of the data in the sense of
  - coverage,
  - completeness,
  - adequacy,
  - exactness etc.

There are two main ways to check the quality of data in a data-set: (1) to check the logical conformity of data inside the data-set; (2) to compare the data with some other data-sets. The first way is quite simply realizable using special software (logic check), but it does not discover all the problems, especially these connected with completeness of the register. The second way is much more effective, but demands special resources and is possible only in the case when the comparable data-sets are harmonized in conceptual and technical level.

The technical problems

- As a rule, the registers have initiated independently from all other data-sets, they have different architecture, ideology and data structure/ description.

Although the registers in general are more modern and flexible data-collections compared with censuses, still they have some problems of flexibility in the sense of development:

When the data amount and description (including definitions and classifications) has been fixed in a register, it cannot be changed easily, when the situation in society changes.

What should be done to follow the IEP?

From here it follows that before using the register data in censuses a big work must be done mainly in the following three directions:

- Preparing the juridical background;
- Harmonization the conceptual and technical structure of different data-sets and registers;
- Creating the common quality standards and procedures for data checking and monitoring for all registers and data-sets.

It is very important that all these steps do not serve the preparations for usage the registers in censuses only, but they are necessary to improve the quality and usefulness of registers as whole.

Chapter 2. Example – the Estonian case

Government wishes the next census should be made using registers

In Estonia since nine censuses have been organized and in 2011 the next, tenth will take place. As the population of Estonia is quite small (1,3 millions) so the statistics – as all administrative undertakings – is quite expensive (per capita). That is why there is quite strong pressure from administration – to make the next census using registers, not in the "traditional way". This situation forced the statisticians to look critically the situation of existing registers and data-bases to assess the possibility of usage them in census.

The list of Estonian registers potentially useable in census

In Estonia the following data-bases exist that might be used in census:

1. Estonian Population Register (EPR), law of EPR from 2000; officially used for registration of statistical events of population – births, deaths, marriages, divorces, also citizenship and migration. The register should have all personal data (including place of birth), current addresses and ID's of close relatives plus some statistical data as education, mother tongue, current social and labor status [4].
2. State Register of Construction Works (SRCW), containing all important data of all buildings and dwellings [5].
3. Estonian Education Register (EER), containing data about all people currently in education system (teaching or learning in all levels), also information about people having graduated some education levels in Estonia during the last ten years; the information about Estonians graduated in foreign countries is incomplete, as depends on the personal initiative [6].
4. Register of Estonian Tax and Custom Board (ETCB), where yearly the data of income and employers are fixed (for all people getting either salary or pension), but since the register does not contain any information about the occupations[7].
5. Registers of Estonian Health Insurance Fund (EHIF), covering all health-insured population (more than 95% of whole population) [8].

Unfortunately, there is no information about the quality of these data-sets. They do not have any quality standards and, seemingly, no regular assessments of data quality have been made.

Besides that, in Statistics Estonia exists the data-base of Census 2000 (C00DB), but it is not useable at the moment due to lack of official permission by Estonian Data Protection Inspectorate.

Juridical restrictions

In Estonia the data protection regulations have been very strong and the Estonian Data Protection Inspectorate (EDPI) has been very active in checking the following the Personal Data Protection Law (2003). This version of PDPL no exceptions exist in personal data processing and

- it is not allowed to use the census-data-base on the level of single records (microdata);
- it is not allowed to link different data-bases/ registers.

The new Data Protection Law, in preparation of which Statistics Estonia participated very actively and that will come into force 01.01.2008 improves the situation somewhat and there are hopes that the C00DB will be open to use it in preparation of the next census.

From here it follows that the systematic assessment of the quality of registers has been since practically impossible. Also the registers' administrators themselves had quite a few motivation to assess the quality of their registers.

Harmonization of registers and identification of records

In all registers recording the ***population*** the records are identified by the ***ID-code*** that is unique and contains also information about sex and age. This identification allows to link these registers easily.

- Here the only problem is that there exist a small number of records in EPR without ID-code. Most of them are very old people, maybe dead or emigrated. But these are still only hypotheses, not checked facts.

Much more problematic is the situation concerning buildings and dwellings, as no ***common address standard*** exists in Estonia and the addresses are different in different registers. The most serious problems arise in the following areas:

- Farms in small villages have been historically identified by farm names (that were, in general, different from owner's name). Part of these names has been forgotten (during collectivization period) and is not in use any more, new buildings have been built in villages etc. At the same time these villages do not have any structure of streets to use the address-standard similar to towns. In fact, these buildings have some technical identification (GP-coordinates and also register codes in cadastral register), but they are not used by population.
- Some street names have different forms (e.g. personal names with and without initials). Some additional problems occur when these personal names have Russian origin (e.g. in Narva and Sillamäe).
- As a result, part of population – these living in small villages without street's structure, but also some other people by different reasons have as living place address only the name of administrative unit (village, town).

From here it follows that the two main registers – EPR and SRCW are not harmonized on the level of dwellings. That makes quite difficult their usage in census.

Two other registers – ETCB and EHIF use the so-called contact (postal or e-mail) address that might be different from living place of a person given in EPR.

It is highly desirable that very soon the common Estonian address standard will be created and established in all registers.

Quality of data of different registers

To assess the quality of Estonian registers some attempts have been made by statisticians especially in the case of Estonian Population Register, but the systematic investigation in this area has been stopped by EDPI. As the EPR has been often used to get the survey samples for Statistics Estonia, several times its quality has been assessed using these samples. The results of these assessments are not very much promising[1], but the quality of data depends strongly on the dimension measured.

- The quality of such data as citizenship, time and place of birth is quite good.
- Some problems are connected with different transcription of names originally not written with Latin letters.
- The main problem is the poor quality of data on living places (errors or gaps in almost 30% of records). Here the reason is that during the period 1992—2004 the registration of living place was not obligatory in Estonia and hence many changes of living place are not reflected in the register. The situation has been somewhat improved during recent years, but the errors made earlier are not corrected, in general, as there are no mechanisms to force the people to register themselves in their real living-place. On the contrary, there are several reasons to register some fictive living-place (better social services and support, better schools, and kindergartens, traveling supports for students etc).
- The data of EPR present some overcoverage compared with the population calculated by Estonian Statistics using C00DB data. Probably, the reason of it is that part of emigrants has not registered their leaving.
- Such data as education and social status are not complete also the classification used for these dimensions in EPR does not coincide with that of censuses.
- Only the legal marital status has fixed (by the standard of EPR). In about 20% of adult persons the marital status has not indicated at all.

Population register does not give immediate information about the families and households, but there are plans to build the Family register on the basis of PR. Still it solves only half of the problem, as it does not give any information about cohabiting couples that are predominant in younger generation in nowadays Estonia [9].

[1] Here the analyses made in Statistics Estonia by Urve Kask, Aira Veelma and others are used., see also [3].

Another problem is that EPR registers only one address for a person, but this seems not to be enough in the case of mobile people who have more than one living place and the “usual” place can be different depending on the definition. This situation has also improved as the second, so called contact address has been added to EPR.

A special project has started to check the quality and coverage of data in the State Register of Construction Works comparing it with the C00DB. As in this comparison no personal data are used, so this comparison is possible without any restrictions.
For this assessment the following design was made. A random sample (consisting of several strata defined by regions and settlement levels) from SRCW register was taken and the same addresses from C00DB will be searched (using special software). In this way the frequencies of following cases will be estimated:

- The addresses of and the whole information in dimensions measured coincided in both data-sets;
- The addresses coincided, but there were differences in other dimensions;
- The addresses did not coincide, but the dwellings were identifiable.
- The address given in SRCW did not exist in C00DB.

As the second step, also the addresses given in C00DB and not existing in SRCW will be identified. Also the reasons of under- and overcoverage of both data-sets will be clarified. The project is in process now.

The quality of data of other registers has not been checked systematically, but the first experience confirms the rather good quality of EER data.

Conclusions

Estonia is not ready for organizing the following census using the registers only. Following the IEP principle, it would be useful

- To use the questionnaires where as much as possible answers are already filled in.
- The information for it should be taken from different sources – EPR, SRCW, C00DB, EER and ETCB data-bases.
- The people should be motivated to check the printed data attentively and correct them, if necessary.

But it is not necessary to make the following census in quite traditional way using interviewers with paper forms and pencils. The following census in Estonia is planned to make in the combined way where CAPI and CAWI will be used. The experience of e-communication is in Estonia quite good, as ETCB has already during five years collected the information mainly using pre-filled forms and since about 80% of all population have filled them in electronically, using identification by bank codes.

The results of census 2011 must be used to make decisions about the quality of registers and to find the possibility to correct them. In some cases it might cause some changes in regulations or laws of registers. After that it might be possible to make the following census using only registers.

References:

1. http://www.statbel.fgov.be/census/links_en.asp Censuses in the world
2. National Practices of UNECE Countries in the 2000 round of Population and Housing Censuses. Draft – April 2006
3. I. Traat, 1996. Leibkonna eelarve uuringu kvaliteedist. ESS Teabevihik 7, 85—93.
4. www.riso.ee
5. www.mk.ee
6. www.ehis.hm.ee
7. www.haigekassa.ee
8. www.emta.ee
9. www. stat.ee