



Economic and Social Council

Distr.
GENERAL

ECE/CES/GE.41/2007/SP/5
29 May 2007

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Group of Experts on Population and Housing Censuses

Tenth session

Astana, 4-6 June 2007

Item 3 (b) of the provisional agenda

CENSUS TECHNOLOGY: RECENT DEVELOPMENTS AND IMPLICATIONS ON CENSUS METHODOLOGY

Supporting Information Sources for Register-Based Census

Submitted by Israel^{*}

This meeting is organised jointly with Eurostat.

Summary

The Bureau of the Conference of European Statisticians (CES), at its meeting held in Washington, D.C. (United States) on 19-20 October 2006, approved the renewed terms of reference for the Steering Group on Population and Housing Censuses and the plan for future CES activities on population and housing censuses. The CES Bureau also agreed that the Steering Group would coordinate the work on the diverse types of meetings. The present paper was prepared for discussion at the Joint UNECE/Eurostat Meeting on Population and Housing Censuses in Astana (Kazakhstan), 4-6 June 2007. The paper provides substantive basis for the discussion in the session of the meeting dedicated to "Census technology: recent developments and implications on census methodology".

^{*} This paper has been prepared by Central Bureau of Statistics, Israel

I. BACKGROUND

1. The next Israeli census will be held on 27 December 2008. It will be an integrated dual census, based primarily on the National Population Register (NPR). The Census will be based on the NPR as the backbone of the census, with two independent samples accompanying it (approximately 15 per cent): a personal interview of all household members in an area sample, to estimate NPR deficits; and a telephone interview of persons sampled from NPR records, to estimate NPR surplus. See a detailed description in <http://www.unece.org/stats/documents/2004/11/censussem/wp.3.e.pdf>.
2. The Israeli NPR was established in November 1948, a few months after the formal establishment of the State of Israel (15 May 1948). The NPR was the main reason for the first census that was conducted jointly by the Central Bureau of Statistics and the Ministry of Interior. The 1948 census was held under a seven-hour curfew in order to ensure that everyone was counted and that no person was counted more than once. The information was based on formal certificates, or by personal declaration if certificates were not available. At the time, a Unique Identification Number (UIN) was assigned to each person present in the household. The information from the questionnaires with the UIN was also recorded on a temporary identification card and entered into the NPR. Children under the age of 18 were registered, and assigned a UIN on one of their guardians' temporary cards. Legal Immigrants and newborns were, and still are, assigned a UIN at the border or at birth. Persons applying for residency status after their arrival, if the application is approved, are given a UIN and their information is entered into the NPR.
3. The NPR includes only residents with a UIN. A person is never removed from the NPR. Deceased and declared emigrants (a resident who applies for a non-residency status) and expelled residents, receive a "NON-ACTIVE" status in the NPR. Records are never removed or deleted from the NPR.
4. Persons not eligible to acquire permanent residency status, or citizenship, are excluded from the NPR.
5. **Register-based census**

The main resources and efforts in a Register-Based Census (RBC) are aimed at overcoming Population Register omissions. There are two types of omissions: population omissions and information omissions. Supporting registers and various administrative sources should help in minimizing both those omissions.
6. **Population omissions**
 - registered population
 - Update of de-facto population to the due date of the census
7. **Information omissions**
 - Address accuracy and update

Anchoring population to buildings and apartments
 Household typology and family structures
 Inherent mistakes
 Update of personal changes

8. **Unavailable information**

Socio-economic characteristics, such as: education, labor force, housing, household equipment, commuting, income, capital.

Population Registers, by nature, are created and maintained for civil purposes. These purposes often differ from conventional census definitions and may impede census estimates.

9. There are several options for overcoming these obstacles:

- a) Creating a “statistical register” which will continuously correct for the civil register flaws. This requires maintaining a parallel Statistical Population Register, with the conventional censal and demographic definitions. The statistical register will need independent updating sources. This possibility relies on the availability of resources, and the cooperation of various governmental and public agencies, as well as population cooperation;
- b) Changing civil definitions and requirements to comply with census definitions – this option is frequently not feasible, and it requires changes in civil laws and maintaining more than one definition of information in the register;
- c) Creating supporting registers – supporting registers are often needed for other purposes. Creating and maintaining registers for the sole need of a census are hard to justify, as they need vast resources and impose a burden on the public;
- d) Modifying and updating the civil register with externally available data sources. Availability and accessibility of current administrative sources are an easy and affordable source for completing and modifying population register omissions. Merging data sources created for various other purposes can create a file that will provide the data needed for a RBC.

10. Administrative sources have their own flaws, as they are created and maintained for administrative purposes. These purposes impose their own definitions, update routines and omissions, which do not necessarily comply with definitions and requirements of censuses. Merging data from different sources will introduce new discrepancies in the merged data set. These discrepancies are solved with a preference procedure for quality, based on the importance of the variable for the data producer. Nevertheless as mentioned, administrative sources are an affordable and accessible source. If their flaws are recognized and corrected for, they can be used as complementary data for a RBC.

II. SUPPORTING SOURCES TO BE USED IN THE 2008 ISRAELI RBC

11. Several administrative files were considered for use in completing and correcting the Population Register, to make an RBC possible. Each one of them was checked for consistency, completeness and its added value. The availability of the UIN, its reliability and update routines, were core factors in the final decision for the inclusion or exclusion of each data source. Some sources were considered for updating addresses; and others for editing or imputing demographic characteristics such as marital status, date of birth and their current residency in Israel. Sources were also considered for their potential in adding information unavailable in the NPR, such as education, income, property ownership, dwelling characteristics, etc. Data sources containing populations excluded from the NPR were a crucial factor in the decision to use their data.

12. Two census pilots carried out in 2004 and 2006/2007 were used to test which data sources would be used to complete and update the data in the NPR.

III. ADMINISTRATIVE DATA SOURCES CONSIDERED

13. Census pilots were carried out in 2002, 2004 and 2006. Following is a summary of the results from the three pilots. Agencies responsible for the potential data sources were approached and asked to deliver the structure and consistence of the requested data, based on the power derived from the “Statistical Order” (the legal term for the Israeli statistical law). Under the law the CBS is entitled to demand any data that is needed for national purposes, as defined by the law (a special appendix is added to the law for each census). Some agencies were reluctant to transmit their data, due to two main problems. First, fear of losing control over data, which is perceived as a source of power and second, potential flaws in the data (not interfering with normal operations). These two obstacles were resolved by negotiations and a commitment by the CBS to reveal potential flaws to the data owner, regardless of their use for census purposes. Some cases required legal negotiation, which prolonged the process. In the latter cases, if the potential added information was negligible, a decision was made to forgo it. Some sources were unacceptable for various reasons.

14. The files that were finally obtained varied, as expected, in their added value. Some of them were based or updated, according to the NPR and therefore had no added information. Others were less accurate than the NPR. Following is a list of the data sources, their potential added information and the final outcomes.

- a) **National Insurance Institute** (social security and national health insurance) collects and maintains data for various purposes: payment of monthly allowances as defined by law, collect taxes (social security and national health insurance), grant special privileges (handicapped, single-parent households, etc.). The information derived from the National Insurance Institute is used for the following modifications and updates in the RBC:
 - i) Inclusion of residents excluded from the Population Register, and exclusion of residents currently in the Population Register, according to data received from other sources;

- ii) Inclusion of non-residents who are eligible for national health insurance, and who have resided legally in Israel for over 12 months;
 - iii) Addition of information regarding monthly income exempted from taxation;
 - iv) Revision of personal status and probably address change, based on changes in eligibility for national health insurance, transfers between health-care providers and changes in eligibility for any allowances;
 - v) Inclusion of information regarding date and type of last update in the database – used only if the last update occurred later than the last update in the NPR.
- b) **Income tax** –Income from work/business and the economic activity branch, of work place /business.
 - c) **Electricity meters** – Address of “registered user” as the potentially most recent address.
 - d) **Vehicle licenses** – The address of the owner of the vehicle registration was used as a potential current address, since vehicle licenses need to be renewed at least once every year.
 - e) **Student enrollment** (all grades – elementary, intermediate, high school) – To add information on actual place of residency.
 - f) **Population registered in local welfare authorities** – Update of household structure and typology. Currently still under investigation.
 - g) **Foreign workers** (entering on a special worker visa) – Adding information excluded from the NPR, and their current addresses.
 - h) **Non-resident students** (entering the country on a student visa)
 - i) **Border control** - Illegal residents (entering on a tourist visa and exceeding their visa time limit).
 - j) **Telephone numbers** - cellular and line phones. Data was used to follow-up on cases, and for verification of actual place of residency.
 - k) **Ministry of Industry, Commerce and Labor** – Additional data on foreign workers

IV. RECORD LINKAGE

15. Before beginning record linkage, checks were conducted to determine, basic validation of their UIN credibility (a legitimate control number) and that they appeared in the NPR. Data sources that passed this initial test were considered for inclusion in the Integrated Administrative File (IAF). The IAF was constructed by linking various micro-data files (individual records). Since data sources are taken from different origins and created for different administrative purposes, the information on each person or household was obtained from administrative sources or personal reports and related to a variety of dates. All sources were matched using record linkage techniques. Record linkage was performed using an algorithm developed specifically for that purpose. The algorithm was carried out in three stages; a deterministic match (exact match), a probabilistic match and a manual match. This algorithm is described fully in a paper at: http://www.fcs.gov/03papers/Yitzkov_AzariaFinal.pdf

16. The backbone source was the NPR. Each file was matched to the NPR in a descending priority order. The outcome was a file containing data from all sources. When variables appeared in more than one source, there was a potential of disagreement on their “true value”. A procedure of grading each variable from each source had to be developed, in order prefer one source over the other. The procedure developed was based on filed work results that verified or disregarded the datum. Some variables, such as address, family relations and names, were kept in the IAF with more than one value, with a code identifying the source.. The census pilots did not give a deterministic preference to any specific source, and validity was related to certain individual characteristics, such as age and marital status. In cases where the datum appeared in more than two sources, the value that appeared in the majority of unrelated sources was considered the “true value”. Raw linked data was edited for several census-critical variables. The editing process was also used to estimate the added value of each data source.

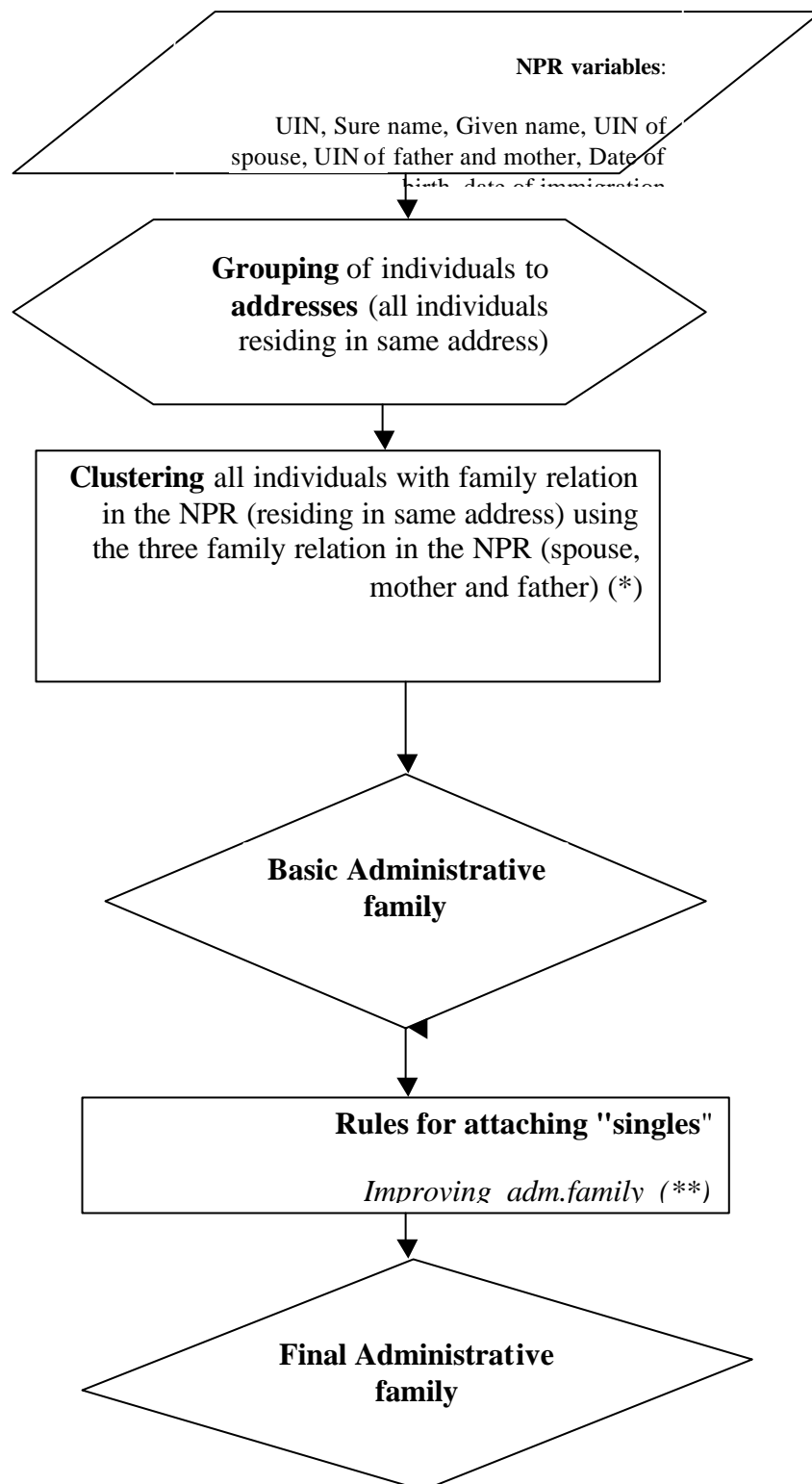
V. CONSTRUCTING "ADMINISTRATIVE HOUSEHOLDS" AND CENSUS ADDRESS-SELECTION RULES

17. The IAF is an individual records file, a single record for each registered person. For the census there is a need for a household identification as well. Administrative households were constructed by an iterative algorithm that uses family relations and address in the IAF (see chart). An administrative family was defined as "a group of people residing in the same address with family relation".

18. "Census address" selection rules:

- a) The NPR address, if all family members resided in the same address and no other source was available;
- b) The address where most of the family members resided if not identical to all family members;
- c) The address of the school children;
- d) The address of the electricity meter;
- e) The address of the car registration.

19. The rules are applied in an iterative procedure. After each step, the administrative family coherence is re-checked and the family is corrected for. With our current experience the procedure converges after a few iterations.

Administrative family construction

(*) Family relations include, direct and indirect relations.

(**) Rules for attaching "singles"

- a) Children (under age 18) living in a different address were merged to their parents.
- b) Single immigrants within the same address (locality, street and number), the same family name and immigrating on the exact date were merged a family.
- c) Married couples with children residing in different addresses were merged to the address of the children.

VI. SUMMARY

20. The NPR was found to be a robust backbone, which made it possible to construct valid individual and household demographic characteristics. Personal demographic information was found to be accurate and updated. The information on family relations proved to be reliable.

21. The administrative sources used, were found to add valuable information, in the creation of administrative families, adding information on income sources and updating addresses.

22. A most useful additions of the supporting administrative sources were the sources that completed the information of residents that are excluded from the NPR and the information, which improved the data on residents not residing currently in the country (absent from the country for over 12 months and residing abroad).

23. The rules for administrative family construction and census address selection will be re-defined as necessary according to the 2006 pilot results.
