UNITED
NATIONS

E

Economic and Social
Council

Distr.
GENERAL

ECE/CES/GE.41/2007/6
22 March 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

Record matching for census purposes in the Netherlands

Submitted by the Netherlands*

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 on request by the Steering Group on Population and Housing Censuses, for presentation and 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".

* The document has been prepared by Statistics Netherlands.

ABSTRACT

1. In the Netherlands, data from many different sources were combined to produce the Census tables of 2001. Since the last Census based on a complete enumeration was held in 1971, the willingness of the population to participate has fallen sharply. Statistics Netherlands found an alternative in the Virtual Census, using available registers and surveys. The table results are not only comparable with the earlier Dutch Censuses but also with those of the other countries in the 2000 Census Round.

2. For the 2001 Census, more detailed information is required than was the case for earlier Census Rounds. The acquired experience in dealing with data of various administrative registers for statistical use enabled Statistics Netherlands to develop a Social Statistical Database (SSD), which contains coherent and detailed demographic and socio-economic statistical information on persons and households. The Population Register forms the backbone of the SSD. The SSD is constructed by micro linking several administrative registers and sample surveys. A micro integration process ensures coherence, consistency and completeness of the SSD data. Sample surveys are still needed for information that is not available from registers.

Keywords: Census; micro integration; micro linking

I. INTRODUCTION

3. The first Census in the Netherlands was held in 1795 for the purpose of establishing voting constituencies. At that time, the united provinces of the Netherlands were still a republic and the borders were different from the current borders. After Napoleon, the Netherlands became a kingdom and once every ten years a census was held. The first Census in the Kingdom of the Netherlands was held in 1829. Before Statistics Netherlands was established, another six Censuses were held in 1839, 1849, 1859, 1869, 1879 and 1889 under the responsibility of the Ministry of the Interior. In 1899, Statistics Netherlands was established and was put directly in charge of the eighth Census. In the 20th century six more traditional Censuses were carried out in 1909, 1920, 1930, 1947, 1960 and 1971. The three most recent Censuses (1981, 1991 and 2001) were not based on a complete enumeration but on registers and surveys available to Statistics Netherlands.

4. Originally, the censuses had two aims. First, they were meant to correct errors in the municipal population registers. Second, they were used to obtain extra information about the socio-economic phenomena in the country. Since the Netherlands conducts a register-based census, the first aim no longer exists. Also, the quality of the central Population Register (PR), which unites all municipality population registers, has improved considerably over time. This is because the incentive for municipalities to keep their population registers up-to-date is the allocation of central government funds among municipalities, which is generally based on the population size according to the local registers. Another reason is that it is very impractical to live in Dutch society without being included in the PR. So both municipalities and citizens have enough incentives to keep the PR of good quality. Recent actions in Rotterdam to improve the quality of the municipal population register for some old quarters prove this statement. The

second aim is still valid and many census results are published in a historical or international context. Currently, census data are popular for comparisons among countries.

5. The first traditional Population Census in the Kingdom of the Netherlands, as laid down by Royal Decree, was held in 1829. It was the start of a hundred and forty year period in Dutch history, during which fourteen traditional censuses were carried out. The final census in this series dates from 1971. Since then a growing distrust in the objectives of a government collecting all sorts of information about its citizens developed. It marked an era in which society became less cooperative, forcing Statistics Netherlands to find alternatives for the traditional population census. Instead of field enumeration, Statistics Netherlands explored administrative registers and sample surveys as new data sources in order to get the necessary census information.

6. For the 1981 and 1991 Census Rounds demographic data were drawn from the Population Register. Data on socio-economic characteristics, such as on labour and education, were provided by the Labour Force Survey. These sources, however, were used separately, which means that no special attention was paid to coherence of the information at the micro level. Moreover, table totals in one source could be different from corresponding totals in the other. To overcome this consistency problem, table results were reweighed to the level of the Population Register totals (for the experiences with the Census of 1981, see Vliegen and Van de Stadt, 1988; for the compilation of the 1991 Census, see Corbey, 1994). The data compiled on 1981 and 1991 were much less detailed than the set of tables of the 2001 Census. Contrary to 1981 and 1991, Statistics Netherlands has published census information for 2001 on the municipal level.

7. For the Census 2001 Programme Statistics Netherlands launched a new approach, which is unique in Europe. One of the most important achievements of the nineties in the area of social statistics is that an increasing amount of socio-economic statistical information can be acquired from administrative registers. Comprehensive and detailed information is now available on employment and social security. By micro linkage and micro integration of demographic and socio-economic data from a wide variety of administrative registers and sample surveys, Statistics Netherlands created a Social Statistical Database (SSD). This SSD contains coherent and detailed information on persons, households, jobs and (social) benefits. Therefore, it is an appropriate data source for the Population Census of 2001. Consistency of sample survey and register (sub)totals is achieved by the method of repeated weighting, a new advanced weighting technique.

8. This paper is largely based on chapters 1 and 13 of Schulte Nordholt et al. (2004). Section II discusses the 2001 Census in some more detail. The method used to compile the 2001 Census is discussed in section III. Combining of data sources, in particular the micro linkage process and the micro integration process is discussed in section IV. Section V goes into the creation of the SSD. Some concluding remarks can be found in section VI.

II. THE 2001 CENSUS

9. In 2003, data were combined to produce the Dutch 2001 Census tables. In the Netherlands, this was done by using data that Statistics Netherlands already had available rather

than by interviewing inhabitants in a complete enumeration. This way, the Dutch taxpayer received a much lower 'census bill'. The costs for a traditional census would be about three hundred million Euros, while the costs using this method are 'only' about three million. The estimate includes the costs for all preparatory work such as developing a new methodology and accompanying software. The costs of the registers are not included (the registers were already there), but the analyses of the results are. Registers are not kept up-to-date for censuses but for other purposes. Saving money on census costs is only possible in countries that have sufficient register information.

10. The 2001 Census relates to forty extensive tables. Twenty-eight are about the Netherlands as a whole, nine are at the COROP level (NUTS 3) and three at municipal level (NUTS 5). The forty tables fall into a number of groups. Eight tables concern housing, two tables concern commuting and the other thirty tables are demographic tables, relating to occupation, level of education and economic activity. Additionally, demographic, housing and labour figures are compiled at sub-city district level for ten large cities that participate in Urban Audit II (Statistics Netherlands, 2003).

11. The Virtual Census in the Netherlands was off to a later start than in other countries where a traditional Census was conducted. It did not make sense to begin the 2001 Census Project until all sources were available; some registers were available relatively late. Nevertheless, the Netherlands was quicker with the compilation of the forty census tables than most of the other countries that participated in the 2000 Census Round. In fact, the Netherlands was one of the first to send the complete set of forty tables to Eurostat, which coordinated the contributions of all European Union (EU) member states, accession countries and European Free Trade Association (EFTA) member states. The Netherlands had the advantage that the incoming census forms did not need to be checked and corrected. However, it must be noted that for some variables only sample information is available, which implies that it was impossible to meet the level of detail required in some Dutch tables.

12. The Nordic countries (Denmark, Finland, Iceland, Norway and Sweden) have currently more variables available in registers than the Netherlands. So the problem of insufficient detail in the outcome does not play a major role there. Moreover, some Nordic countries conducted a (limited) enumeration for variables missing in the registers. Most of the other countries are in a similar position as the Netherlands where some variables relevant for the census can be found in registers, while other variables are available on a sample basis only. That's why much interest exists in the Dutch approach to combine registers and surveys and to use modern statistical techniques and software to compile the tables. It is, of course, crucial that statistical bureaus be able to make use of registers that are relevant for the census. For Statistics Netherlands, this possibility was strengthened in the new statistical law of 2003. Nevertheless, in the years to come Statistics Netherlands will have to establish good contacts with register holders. Timely deliveries with relevant variables for Statistics Netherlands are essential for statistical production.

13. More than fifty countries in the United Nations Economic Commission for Europe region participated in the 2000 Census Round. Many countries chose a day in 2001 as their reference day, although they chose many different days. As it takes a long time before all countries finish the tables required by the international organisations, the Netherlands took the initiative to make some simple comparisons among nine European countries that were relatively quick in

compiling the set of tables for Eurostat and that were willing to join the comparison analyses (see Chapter 12 in Schulte Nordholt et al., 2004). The results of the Dutch 2001 Census were also compared to earlier Dutch Censuses (Schulte Nordholt, 2005).

III. METHOD OF COMPILING OF THE 2001 CENSUS

14. The current virtual census relates to 2001. The backbone of the census is the central Population Register (PR), which is the combination of all municipal population registers. The Population Register (PR) contains demographic information on every inhabitant of the Netherlands (Prins, 2000).

15. Even though the PR seeks to optimally record every person in the population, it is by no means perfect. People may move to live elsewhere and 'forget' to notify the authorities. Therefore, municipal population registers are not always up-to-date. Another example of improper registration in the PR is when two persons are registered at separate addresses, but actually live together. They have a financial incentive to be registered at different addresses if one person is employed and the other is on welfare. This is because the person receiving benefits might loose them when the social security agency finds out they are living together.

16. An important population group that the PR misses are the people who live in the country without the authorities' knowledge, many staying illegally. This population group is not present in the census population. Illegal residents pose a problem for statistical offices because, on the one hand, they participate in the economy and as such they are included in economic statistics. On the other hand, they are not covered by demographic statistics. It is very unlikely that they would be enumerated in a traditional census though. Statistics Netherlands has made an attempt to estimate the size of the illegal population, but since there is hardly any information this proved to be very difficult. The official estimate of the number of illegal residents on 1 January 2001 by Statistics Netherlands is one with a wide margin: between 46 thousand and 116 thousand people (Hoogteijling, 2002).

17. A number of integrated surveys and registers were linked to the PR. For this linking process only exact matches are used based on the unique so-called Social security and Fiscal (SoFi) number. The integrated system is called the Social Statistical Database (SSD) system. It is developed originally to conduct virtual censuses, but now it is also used for many social statistics.

18. PR data of 1 January 2001 were used as the basis for the set of tables. Different variables, such as occupation and level of education, were obtained from the Labour Force Survey (LFS). The variable job size was obtained from the large Survey on Employment and Earnings (SEE). To obtain sufficient records, information on persons from the LFS 2000 and the LFS 2001 was combined. For the housing tables, we used PR data of 1 January 2001, the Housing Register 2001 and the Survey on Housing Conditions (SHC) 2000.

19. Some variables of the PR and SSD datasets are available on an integral basis. Examples are age, sex, marital and employment status. Survey variables are only available for a part of the population. Examples are the highest level of education attained (LFS) and whether someone rents or owns the property they live in (SHC).

20. To be able to estimate every table as accurately as possible, each estimate is based on the largest possible number of records. Tables that contain register variables only are counted from the registers. Tables that contain at least one variable from a survey are estimated from the largest possible combination of registers and surveys.

21. We guaranteed consistency among the tables by using the technique of repeated weighting. It generates a new set of weights for each estimated table and is based on the repeated application of the regression estimator. When using repeated weighting, the weights of the records in the microdata are adapted in such a way that a new table estimate is consistent with all earlier table estimates.

22. The figures of the 2001 Census relate to persons living in the Netherlands on 1 January 2001 (counting unit persons). The persons who were living in the Netherlands at the beginning of that day according to the PR were 'counted' in the Virtual Census. Most of the Dutch population lives in private households, the remainder being part of institutional households. The number of employees in the tables relates to the end of the year 2000 for which 22 December 2000 was used as reference date to fix the number of jobs of employees in the Netherlands. It was impossible to have a reference day in 2001 for the number of employees since the SSD datasets 2001 were not available in time to use in the 2001 Census. The SSD data used registers' information on the jobs of employees. If an employee holds several jobs at the same time, he or she can appear several times in the employee register. The features of the main job are used in the set of tables. The main job of an employee has been defined as the job with the highest gross wage for the social insurances.

IV. COMBINING DATA SOURCES: MICRO LINKAGE AND MICRO INTEGRATION

A. Micro linkage

23. Most of the present administrative registers are provided with a unique linkage key. It is the so-called social security and fiscal number (SoFi-number), a personal identifier for every (registered) Dutch inhabitant and those abroad who receive an income from the Netherlands and have to pay tax over it to the Dutch fiscal authorities.

24. To prevent misuse of the SoFi-number, Statistics Netherlands recodes it for statistical processing into a so-called Record Identification Number (RIN-person). Personal identifiers, such as date of birth and address, are replaced by age at the reference date and RIN-address. This is all done in accordance with regulations of the Dutch Data Protection Authority to protect the privacy of the citizens.

25. Since the SoFi-number is in use by social security administrations and tax authorities, one may expect it to be of excellent quality. A limited amount of SoFi-numbers may be registered with incorrect values in the data files, in which case linkage with other files is doomed to fail. However, in general, the percentage of matches is close to one hundred percent. Abuse of SoFi-numbers, for example by illegal workers, may occur in some cases, which results in a false match. Sometimes there are indications of a mismatch. An example of this is when the jobs register and the PR are linked and the worker turns out to be an infant. Another example is, when

the FiBase (fiscal administration) shows an unusually high income for a worker, when it is in fact the sum of the incomes of all people using the same SoFi-number.

26. All social statistics data files can be linked to the PR. In practice this means that these data files are all indirectly linked to each other via the PR. Therefore the PR can be considered the backbone in the set of social data sources. When linking the PR and the jobs register, or the PR and a register of social benefits, it is a linkage between different statistical units (persons, jobs, benefits). In that case multiple linkage relationships can exist because someone can have more than one job or can benefit from several social benefits.

27. In household sample surveys, like the LFS, records do not have a SoFi-number. For those surveys an alternative linkage key is used, which is often built up by a combination of the following personal identifiers:

(a) sex;
(b) date of birth;
(c) address[1].

28. This sort of linkage key will usually be successful in distinguishing people. However, it is not a 100 percent unique combination of identifiers. Linking may result in a mismatch in the case of twins of the same sex. False matches may also occur when part of the date of birth or the postal code and house number is unknown or wrong. Another drawback is that the linkage key is not person but address related, which may cause linkage problems if someone has recently moved. When linking the PR and the LFS with this alternative key, and tolerating a variation between sources in a maximum of one of the variables sex, year of birth, month of birth or day of birth, the result is that close to 100 percent of the LFS records will be linked.

29. In its linkage strategy, Statistics Netherlands tries to maximize the number of matches and to minimize the number of mismatches. So, in order to achieve a higher linkage rate, more efforts are made to link the remaining unlinked records by means of different variants of the linkage key. For example, leaving out the house number and tolerating variations in the numeric characters of the postal code. To keep the probability of a mismatch as small as possible, some 'safety' devices are built in the linkage process. This last linking attempt accomplishes an extra one percent matches.

30. In the end about two to three percent of the LFS records could not be linked to the PR. All together this is a good result, but selectivity in the micro linkage process is not to be ruled out. If the unlinked records belong to a selective subpopulation, then estimates based on the linked records may be biased, because they do not represent the total population. Analysis in the past has indicated that the young people, in the 15-24 age bracket, show a lower linkage rate in household sample surveys than other age groups. The reason for this is that they move more frequently, therefore they are often registered at the wrong address. The linking rate for persons living in the four large cities Amsterdam, Rotterdam, The Hague and Utrecht is lower than for

[1] In fact, the combination of a postal code (mostly related to the street) and house number is used as substitute for the address. The postal code in the Netherlands consists of four figures, followed by two letters.

persons living elsewhere. Ethnic minorities also have a lower linkage probability, among other things because their date of birth is often less well registered (Arts et al., 2000).

31. Nowadays, the PR is serving as a sampling frame for the LFS. Therefore, the matching rate is almost 100 percent, and no more linkage selectivity problems occur.

B. Micro integration

32. Successfully linking the PR with all the other data sources mentioned, makes much more coherent information on the various demographic and socio-economic aspects of each individual's life available. One has to keep in mind, however, that some sources are more reliable than others. Some sources have a better coverage than others, and there may even be conflicting information between sources. So, it is important to recognize the strong and weak points of all the data sources used.

33. Since there are differences between sources, we need a micro integration process to check data and adjust incorrect data. It is believed that integrated data will provide far more reliable results, because they are based on an optimal amount of information. Also the coverage of (sub) populations will be better because when data are missing in one source we can use another source. Another advantage of integration is that users of statistical information will get one figure on each social phenomenon, instead of a confusing number of different figures depending on what source has been used.

34. During the micro integration of the data sources the following steps have to be taken (Van der Laan, 2000):

(a) harmonisation of statistical units;
(b) harmonisation of reference periods;
(c) completion of populations (coverage);
(d) harmonisation of variables, in case of differences in definition;
(e) harmonisation of classifications;
(f) adjustment for measurement errors, when corresponding variables still do not have the same value after harmonisation for differences in definitions;
(g) imputations in the case of item non-response;
(h) derivation of (new) variables; creation of variables out of different data sources;
(i) checks for overall consistency.

All steps are controlled by a set of integration rules and fully automated.

35. Now an example follows of how micro integration works is the case in which data from the jobs register are confronted with data from the register of benefits. Both jobs and benefits are registered at volume base, which means that information on their state is stored at any moment in the year instead of at one reference day. Analysts of the jobs register know that the commencing date and the termination date of a job are not registered very accurately. It is important though to know whether or not there is a job at the reference date, in other words whether or not the person is an employee. With the help of the register of benefits it is sometimes possible to define the job period more accurately.

36. Suppose that someone becomes unemployed at the end of November and gets unemployment benefits from the beginning of December. The jobs register may indicate that this person has lost the job at the end of the year, perhaps due to administrative delay or because of payments after job termination. The registration of benefits is believed to be more accurate. When confronting these facts the 'integrator' could decide to change the date of termination of the job to the end of November, because it is unlikely that the person simultaneously had a job and benefits in December. Such decisions are made with the utmost care. As soon as there are convincing counter indications of other jobs register variables, indicating that the job was still there in December, the termination date will in general not be adjusted.

V. THE SOCIAL STATISTICAL DATABASE (SSD)

37. The micro linkage and micro integration process of all the available data sources result in the end in the Social Statistical Database (SSD), a whole set of integrated microdata files in their definitive stage. The SSD contains coherent and detailed demographic and socio-economic statistical information on persons, households, jobs and (social) benefits. A major part of the statistical information is available on volume base. An extensive discussion on the SSD can be found in Arts and Hoogteijling (2002).

38. In trying to imagine what the SSD looks like, one should not think of a large-scale file with millions of records and thousands of variables. It would be very inefficient to store the integrated data as such. Furthermore, the issue of data protection prevents Statistics Netherlands from keeping so much information together. Instead, all the integrated files in their final stage are kept separately. There is just one combining element, which is the linkage key RIN-person, present in every integrated file. So, whenever users demand a selection of variables out of the SSD set, only the files with the variables demanded will be supplied. These can easily be extracted from the set and linked by means of the linkage key.

VI. CONCLUDING REMARKS

39. Statistics Netherlands has innovated its methods of data collecting and data processing for the compilation of the Census Table Programme 2001. The most important elements in the new approach are the use of a combination of administrative registers and sample surveys as data sources, and the application of repeated weighting, a new methodology to estimate numerically consistent tables from this data source. The result is called a virtual census, because the results for some characteristics of the population are based on estimates instead of enumeration.

40. The new way of producing census tables proved to be a succesful and much cheaper alternative for the costly census projects of the past. No special effort had to be made to collect data, as the data sources used for the 2001 Census were already part of the regular statistical programme of Statistics Netherlands. Most data came from registers, and only some supplementary information was needed from sample surveys. This means that the implementation of the 2001 Census Programme did not cause any extra response burden. Moreover, the data processing time for the 2001 Census was just a fraction of what it would have been in a traditional census.

41. One disadvantage of the new approach is that the sample size of the surveys used is not always sufficient to guarantee reliable estimates for small subpopulations, such as municipalities. A traditional census would not have such a problem.

42. When one compares the present way of compiling census tables with those used in the Censuses of 1981 and 1991, the Census Programme 2001 has required more production time but the estimates in cross tabulations of register and survey information are more accurate. First, because the statistical information has gained much more coherence because of combining the data sources. Second, because more auxiliary information could be used in the estimation methods than in the past since more registers are available. Third, because much effort has been made to achieve overall numerical consistency.

Bibliography

Arts, C.H. and E.M.J. Hoogteijling. 2002. 'The Social Statistical Database of 1998 and 1999'. *Monthly Bulletin of Socio-economic Statistics.* Vol. 2002/12 (December 2002), pp. 13-21, 2002. [in Dutch]

Corbey, P.I., 1994. 'Exit the population census'. *Netherlands Official Statistics*, Vol. 9 (Summer 1994): pp. 41-44, 1994.

Hoogteijling, E.M.J. 'Illegal people in the Netherlands'. *Monthly Bulletin of Population Statistics.* Vol. 2002/03 (March 2002), page 21. [in Dutch]

Laan, P. van der, 2000. 'Integrating Administrative Registers and Household Surveys'. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, *Integrating Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker, pp. 7-15.

Prins, C.J.M., 2000. 'Dutch population statistics based on population register data'. *Monthly Bulletin of Population Statistics.* Vol. 2000/02 (February 2000), pp. 9-15.

Schulte Nordholt, E., M. Hartgers and R. Gircour (Eds.), 2004. 'The Dutch Virtual Census of 2001, Analysis and Methodology'. Statistics Netherlands, Voorburg / Heerlen, July, 2004. http://www.cbs.nl/NR/rdonlyres/D1716A60-0D13-4281-BED6-3607514888AD/0/b572001.pdf.

Schulte Nordholt, E., 2005. 'The Dutch virtual Census 2001: A new approach by combining different sources'. *Statistical Journal of the United Nations Economic Commission for Europe*, Volume 22, Number 1, 2005, pp. 25-37.

Statistics Netherlands, 2003. 'Urban Audit II, the implementation in the Netherlands'. Report, BPA no. 2192-03-SAV/II, Statistics Netherlands, Voorburg. http://www.cbs.nl/NR/rdonlyres/8C6E4C9D-4338-4E32-848B-8D43B9B3242D/0/urbanauditIINetherlands.pdf.

Vliegen, J.M. and H. van de Stadt, 1988. 'Is a Census still necessary? Experiences and alternatives'. *Netherlands Official Statistics,* Volume 3, No. 3, pp. 27-34.

*****