Distr.: General
27 March 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
session 4: Methodology for estimating information missing in registers

Methodology used for estimating Census tables based on incomplete information

Note by Statistics Netherlands [1]

Summary

Data from many different sources are combined to produce Dutch census tables. 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 a virtual census, using available registers and surveys. A virtual census is cheaper, comparable to earlier Dutch censuses, and more socially acceptable. The Netherlands takes up a unique position in the European Census Round of 2011: the Dutch approach is register-based and census questionnaires no longer exist; missing information is estimated from already existing surveys and the methodology of repeated weighting is applied as key method for the production of the set of census tables. The table results are not only comparable with the earlier Dutch censuses but also with those of the other European countries that take part in the 2011 Census Round.

I. Introduction

1. In the Netherlands two years after a census year microdata are combined to produce the Dutch census tables. In the Netherlands this was not done by interviewing inhabitants in a complete enumeration, but by using data that Statistics Netherlands already has available.

[1] Prepared by Eric Schulte Nordholt.

This way, the Dutch tax payer gets a much lower census bill. The costs for a traditional census would be a few hundred million euros, while the costs made currently are ‘only’ a few million. This estimate includes the costs for all preparatory work such as developing, updating and testing the methodology and accompanying software. The costs of the registers are not included, but the analyses of the results are. Registers are not kept up-to-date for censuses but for other purposes. Countries conducting a traditional census often justify the huge census costs by pointing out the enormous implications of the census results for the distribution of government money among regions. Moreover, a virtual census would be impossible in many countries because of the lack of sufficient register data.

2. Except the financial aspect, other important differences exist between a traditional census and the virtual census conducted in the Netherlands. In spite of the mandatory character of a traditional census, a certain part of the population will not participate (unit non-response) and the part that does participate will not answer some questions (item non-response). Correcting non-response by weighting and imputation techniques is well worth trying. A well-known problem with traditional censuses is that participation is limited and selective. Traditional correction methods fall short of the need to be able to publish reliable results. The last traditional census in the Netherlands (in 1971) met with much privacy objections against the collection of integral information about the population living in the Netherlands. This increased the non-response problem and the expectation was that non-response would be even higher if another traditional census were held in the Netherlands (Corbey, 1994). There are almost no objections to a virtual census and the non-response problem only plays a role in the surveys of which the data are used. If non-response can be corrected in a survey, it will certainly be possible to correct for the selectivity of that survey in the census where it is used.

3. The virtual census in the Netherlands is off to a later start than in other countries where a traditional census is conducted. It does not make sense to really start the main phase of the Census Project until all sources are available; some registers are available relatively late. Nevertheless, the Netherlands will probably be quicker with the compilation of the census tables than most of the other countries that participate in the European Census Round. In fact, the Netherlands has one of the shortest production times before the complete set of tables is provided to Eurostat, which co-ordinates the contributions of all European Union (EU) member states, accession countries and European Free Trade Association (EFTA) member states. The Netherlands has the advantage that no incoming census forms need to be checked and corrected. However, one must realise that for some variables only sample information is available, which implies that it is impossible to meet the level of detail required in some of the census tables.

4. Currently, in the Netherlands the advantages of a virtual census in cost and non-response problems amply make up for the loss of some detail compared to a traditional census. Moreover, not all information required will always be available for the users in traditional censuses. This is because traditional correction methods such as weighting and imputation sometimes do not correct for limited and selective participation. This means that no reliable results can be published for some of the cells in the set of tables.

5. Four Nordic countries (Denmark, Finland, Norway and Sweden) have more variables available in registers than the Netherlands. So the problem of insufficient detail in the outcome does not play a major role there. Also in Austria and Slovenia most census variables are available in 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 is why much interest exists in the Dutch approach to combine registers and surveys and to use modern statistical techniques and accompanying software to compile the tables. It is of course crucial that statistical bureaus are permitted to make use of registers that are relevant for the Census.

For Statistics Netherlands this was laid down in the new statistical law that came into force in 2004. Nevertheless, in the coming years Statistics Netherlands will have to continue the good contacts with register holders established over the last fifteen years. Timely deliveries with relevant variables for Statistics Netherlands are crucial for statistical production.

6. In the Netherlands the Census of 1971 was the last census using questionnaires. The Dutch parliament decided in the seventies of last century that no more traditional censuses will be held in the Netherlands. At the same time there was still a huge interest in census results. The census data compiled on 1981 and 1991 were much less detailed than those for 1971 and the set of tables of the Virtual Census of 2001. To be able to produce more detailed tables for the 1981 Census for the Labour Force Survey of that year a higher sampling fraction than normal was drawn. Later on, this oversampling of the LFS for census purposes was no longer executed and censuses made use of regular LFS data only. The 1991 Dutch census was largely based on a register count of the population in combination with the Labour Force Survey 1991 and the Housing Demand Survey 1989/1990. In 1991 the Census Act was rescinded, officially cancelling Statistics Netherlands obligation to hold a census once every ten years (Corbey, 1994). There was no European obligation to supply census data on 2001, but it was considered inconceivable that the Netherlands would not compile census data for the international organisations just like all other European countries did. Contrary to 1981 and 1991, Statistics Netherlands has published census information over 2001 on the municipal level.

7. The reason why Statistics Netherlands has compiled the set of tables of the 2001 Census is based on a gentlemen’s agreement with Eurostat that made it possible to compare the 2001 Census tables among European countries. The results of the Dutch 2001 Census were also compared to earlier Dutch censuses. Such work has been carried out in the past as well. A general feeling after the 2001 Census in Europe was that a gentlemen’s agreement was not enough to continue the ten-yearly censuses in Europe. This was the reason to give the 2011 Census Round a broader basis with four Regulations (European Commission, 2008, 2009, 2010a and 2010b). With these four regulations, the population definitions, census variables and their categories, census hypercubes (high-dimensional tables) and metadata are harmonised within the EU. Moreover, the technical format in which the data have to be delivered has been specified and all countries will produce a quality report in which the methodology used is described.

II. Method of compiling

8. Statistics Netherlands has conducted virtual censuses since the 1981 Census. The backbone of these censuses is the central Population Register (PR), which is the combination of all municipal population registers. The census results relate to persons living in the Netherlands on Census day (counting unit persons). The persons who are living in the Netherlands at the beginning of that day according to the PR are ‘counted’ in the virtual census. Most people in the Netherlands live in private households, the others are part of institutional households. For the 2001 and 2011 Censuses PR data of 1 January (Census day) are used as the basis for the set of tables. The set of tables focuses on frequency counts and not on quantitative information.

9. In the nineties of last century the Social Statistical Database (SSD) system was set up at Statistics Netherlands. The SSD includes integrated microdata on employees and self-employed. The SSD is a set of integrated microdata files with coherent and detailed demographic and socio-economic data on persons, households, jobs, benefits and dwellings. Because of the micro-integration process executed before the data are stored, the SSD contains no remaining internal conflicting information. For the 2001 Census some variables (e.g. job size) were obtained from the large Survey on Employment and Earnings

(SEE). This survey stopped in 2006 and since then Statistics Netherlands is relying more on fiscal and social security data that have been included in the SSD. The number of employees in the tables relates to the end of the year before the census year. To define the end of the year a day in December is used as reference date to fix the number of jobs of employees in the Netherlands. It is impossible to have a reference day in the census year for the number of employees, since the SSD datasets of the census year are not available on time to use in the census. The SSD data used register 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. In the set of tables the features of the main job are used, in which the main job of an employee has been defined as the job with the highest gross wage for the social insurances.

10. Different variables, such as occupation and level of education, are obtained from the Labour Force Survey (LFS). For the variable level of education in the 2011 Census some register information (based on examination results) is available in addition to the LFS. To obtain sufficient records, information on persons is combined from the LFS of the year before the Census year and the LFS of the Census year. This way, it is assumed that the scores on the variables in the LFS are stable in the two year period around Census Day. Between the 2001 and 2011 Censuses the LFS has changed into a panel survey. Also the new panel character can now be used: data from different waves are available and the data closest to Census day can be used to compile the tables.

11. For the housing tables in the 2001 Census we used PR data of 1 January 2001, the Housing Register 2001 and the Survey on Housing Conditions (SHC) 2000. For the 2011 Census housing information is obtained from the housing register and new sources like the Basic registrations on Addresses and Buildings (BAB). For the 2001 tables on commuting we used the PR data, the SEE 2000 and the SSD datasets of 2000. For the 2011 tables on commuting we will rely on the PR for the place of residence and on SSD data for the place of work.

12. Register 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 and whether someone rents or owns the property they live in. Austria, Denmark, Finland, Norway, Slovenia and Sweden have indicated us recently that they are planning to conduct a fully register-based 2011 Census. Although they all miss some information, the amount of information they miss is more limited than in the Netherlands. Therefore, some imputation methods are applied to compensate for the relevant missing information. For the Netherlands this is not a viable strategy since we miss somewhat more information than the six countries conducting a fully register-based 2011 Census.

13. Another issue for register-based countries is the difference between the official definitions of the Census 2011 variables and the variables available in registers. An important example is the number of unemployed people. In the International Labour Organisation (ILO) definition it is stated that persons can only be considered unemployed if they are actively looking for work. From registers it can normally not be concluded whether a person who receives a social security or unemployment benefit is indeed actively looking for work. Moreover, in a register-based approach one misses those unemployed who do not receive a social security benefit (e.g. because other household members earned sufficient income). For the 2001 Census the Netherlands used the official ILO definition as requested and raised the sample results to the population totals. With the more detailed hypercubes for the 2011 Round Statistics Netherlands seriously considers an unemployment definition based on registers only. That way, the level of detail required in the tables on current activity status (for which unemployment is one of the categories) can be reached. If Statistics Netherlands indeed decides so, the unemployment information in the census

hypercubes of Statistics Netherlands will be well comparable with the other register-based countries, but the comparison with the other European countries will become more problematic.

III. Repeated weighting

14. One may wonder why simply applying mass imputation (filling in valid values for all missing scores) is not taken into account to overcome the missing data problems. An important advantage of mass imputation is that once the records are imputed, any user will be able to reproduce results when using the same imputed file. However, mass imputation is not a viable strategy for raising survey outcomes to population totals. There are not enough degrees of freedom to sustain a sufficiently rich imputation model accounting for all significant data patterns between sample and register variables. Only if the interest is in totals of subsets of the population defined by the explanatory variables in the model, the imputation approach leads to approximately design-unbiased and hence reliable estimates (at least if the variances are reasonably small) (Kroese and Renssen, 2000).

15. The key method for the 2011 Census in the Netherlands is the technique of repeated weighting (RW) described by Houbiers et al. (2003) and Houbiers (2004). The aim of repeated weighting is to cut out numerical inconsistencies among table estimates from different sources. It is based on the repeated application of the regression estimator and generates a new set of weights for each table that is estimated. Let *y* be a variable of which the population parameter - either total or mean - ought to be obtained for a table through a set of explanatory variables *x* from a register. The linear regression estimator of the population mean for *y* is defined by

$$\hat{\bar{Y}}_{REG} = \hat{\bar{Y}}_d + b'_s\left(\bar{X}_p - \hat{\bar{X}}_d\right); \qquad b_s = \left(X'_s D_s X_s\right)^{-1} X'_s D_s y_s ,$$

where $\bar{X}_p$ and $\bar{Y}_p$ are the population means of *x* and *y*, respectively while $\hat{\bar{X}}_d$ and $\hat{\bar{Y}}_d$ are their estimates based on the design weights and b_s is the estimated vector of regression coefficients. X_s is the matrix of sample observations on the *x*-variables and y_s is the vector of observations on the variable *y*. D_s is the diagonal matrix with design weights. Instead of these traditional regression estimators, the repeated weighting procedure uses a set of coefficients in the form

$$b_w = \left(Z'_s W_s Z_s\right)^{-1} Z'_s W_s y_s ,$$

where Z_s is the matrix of sample observations on the variables in the margins of the table with variable *y* and W_s is the diagonal matrix with regression weights of the last weighting step to calculate the regression estimator. The means of the marginal variables *z* have been estimated already in an earlier table or are known from a register. Denoting these estimates or register counts by $\hat{\bar{Z}}_{RW}$, the repeated weighting estimator of $\bar{Y}$ is defined by

$$\hat{\bar{Y}}_{RW} = \hat{\bar{Y}}_{REG} + b_w\left(\hat{\bar{Z}}_{RW} - \hat{\bar{Z}}_{REG}\right).$$

It can be shown that the weights of the records in the microdata are adapted in such a way that the new table estimate is consistent with all earlier table estimates (Knottnerus and Van Duin, 2006).

16. To apply the technique of repeated weighting we use the latest version (2.0) of the software package VRD developed by Statistics Netherlands. The letters VRD stand for Vullen (Filling) Reference Database and the aim of the application is to fill and manage the reference database. The main functions of VRD are the estimating of tables via repeated weighting, adding these tables to the reference database, and withdrawing aggregates from the reference database. Under the condition of small, independent samples, the variances of the table values can also be estimated. The estimating of the tables does not occur in VRD itself, but takes place in Bascula automatically without the VRD-user seeing this explicitly. Estimating the tables and the variances can be done in batch mode or interactively.

17. To be able to estimate every table as accurately as possible, every 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. The combination of registers and surveys form blocks from which the census tables are estimated.

18. For the blocks that are compiled on the basis of survey data weights have to be determined to produce estimates for the complete population. These weights depend on:

- the precise composition of the block concerned (one or more surveys);

- the design of the survey(s);
- the non-response correction of the survey(s);
- the reduction of the variance by means of auxiliary information;
- the reaching of consistency.

19. Complete consistency is not always possible, for example if too many restrictions are imposed. In some cases complete consistency is possible, but it leads to a very large variation in the weights and thus increases variance drastically. In those cases it is better to restrict the detail that is published.

20. In compiling the census tables we adapt the weights of the blocks at every VRD turn by means of all relevant register counts and the tables estimated earlier from the blocks. This way, all tables are mutually consistent. Every table has to be calculated from the largest block from which the table can be determined. If all tables are estimated this way with the correct weights, the tables' results are mutually consistent. By starting every time from the largest block, the most detailed possible census tables are achieved.

21. The results of five simulation studies testing various aspects of repeated weighting can be found in Van Duin and Snijders (2003). Knottnerus and Van Duin (2006) give the variance formulae for the repeated weighting (RW) estimator, and test RW estimators under various conditions. How the estimated margins of the estimation results are used to decide which cells can be published and which cells have to be suppressed is explained in the next section.

22. It remains an open issue how small areas can be estimated in case of no available register information. Current research is conducted on complementary methods to minimise the empty cells problem. Zeroes as cell values based on sample data do not necessarily mean that no data exist at the population level. However, with standard weighting techniques a sample zero cannot be raised to a positive cell value in the population.

23. An interesting option is to use small area estimation techniques to estimate the cell values that could not be estimated adequately. The estimation of small areas, that is to say, to get a valid and efficient estimation of population parameters for sub national domains, both geographically-based domains or categories in classifications at a very disaggregate level, is a task that can be properly done by making use of administrative registers. For efficient small area estimation, records from an administrative source can play the role of auxiliary information. A theoretical framework for small area estimation can be found in

Rao (2003). The Office for National Statistics (ONS) in the United Kingdom (UK) studied the application of this technique in the context of its Neighbourhood Statistics Programme. This is a major initiative to bring together and make widely available statistics on a small area level. In each case of implementation of indirect small area estimates particular attention was paid to model specification. Some experimental synthetic estimates were published in the UK and others are undergoing a process of evaluation. Possibly, the techniques of repeated weighting and small area estimation can be combined in the 2011 Census Round. However, the open issue is how to keep consistency of the set of tables.

24. Another interesting option is macro-integration (Mushkudiani, Daalmans and Pannekoek, 2012). Macro-integration is widely used for the reconciliation of macro figures, usually in the form of large multi-dimensional tabulations, obtained from different sources. Traditionally these techniques have been extensively applied in the area of macro-economics, especially in the compilation of the National Accounts. Methods for macro-integration have developed over the years and have become very versatile techniques for solving integration of data from different sources at a macro level. A current research project at Statistics Netherlands tries to apply macro-integration techniques in the reconciliation of tables of the virtual census.

IV. How to publish the tables?

25. The Dutch census is compiled partly on the basis of sample data. Therefore, margins of inaccuracy have to be taken into account for some census results. Because of the reliability of the results, rules of thumb are being applied for cell values that are based on a sample from the census population. The exact margins of inaccuracy cannot be given because blocks are composed from the surveys and because of the complex design of these surveys. The rules of thumb are deduced on the basis of the assumptions that the two LFS datasets (for the census year and the year before the census year) form one sample and that the ‘inclusion probabilities’ for this sample are given by the block weights of the LFS block. The rules of thumb in the 2001 Census for records of observations from the LFS run as follows:

- Table cells based on less than 10 persons are always suppressed.
- Table cells based on 25 or more persons are always published.
- Table cells based on 10–24 persons are only published if they form a part of a breakdown (by age or sex), in which no cells based on less than 10 persons occur, and at least 50 percent of the cells in the breakdown have more than 25 persons. The threshold of 25 persons corresponds to an estimated relative inaccuracy of at most 20 percent (i.e. the estimated margins amount to 40 percent at most).

26. The rules of thumb for records from the SHC are of the same form. However, somewhat higher threshold values are applied because of the fact that the sample size of the SHC is somewhat more limited than the one of the LFS. For table cells with households or dwellings as counting unit, analogous rules of thumb are applied for the Dutch Census of 2001.

27. It is to be expected that for the 2011 Census similar rules of thumb can be achieved. The confidentiality problems will largely be solved when all decisions whether to suppress a table cell or not are made according to these rules of thumb. Most census variables are not sensitive and most census tables are frequency tables where confidentiality issues are less harmful for the detail of the output than for quantitative tables. Different options how to protect all output of the 2011 Census Round properly will be discussed at the ESSnet Workshop on Statistical Disclosure Control (SDC) of Census data in Luxembourg on 19-20

April 2012. This Workshop is part of the ESSnet on common tools and harmonised methodology for SDC in the European Statistical System (ESS).

28. All EU member states have to conduct a 2011 Census. This was for most National Statistical Institutes a major operation that involved a lot of work and high costs. All countries have to validate and protect the census output in the form of hypercubes. Even the formats used for the data will differ from country to country. However, in the end all data have to be transformed to SDMX format and offered to Eurostat. Eurostat produced the DSDs (Data Structure Definitions) for the delivery and made them available to all member states. Individual country checks on confidentiality are evaluated within the context of this ESSnet. Although the actual delivery deadline of all hypercubes to Eurostat is only in 2014, before that deadline all protection measures should be taken.

29. It will be very profitable if European countries exchange experiences and learn from each others census confidentiality approach. Now we face the risk that many countries stay at the safe side and protect too much information. This could hamper the calculation of European totals. Also the situation where all countries suppress different cells in the hypercubes will lead to the problem that some totals cannot be calculated at the European level. The ESSnet Workshop in April 2012 is also held to solve this problem. For this workshop key note speakers from different countries have been invited to give their view on the European Census confidentiality problem. Then an open discussion will be organised with an employee from Eurostat as moderator. This workshop will be conclusive, recommending certain confidentiality methods for the European Census. By this activity an enormous step forward will be made in harmonising the protected output between European countries. This way, more and better comparable Census output can be produced with minimum information loss.

V. Conclusions

30. The virtual census has proved to be a successful concept in the Netherlands. It has many advantages compared to traditional censuses. The census costs are now considerably lower and nevertheless data on the Netherlands become available that can be compared to results of earlier Dutch censuses and to the results of other countries that take part in the same Census Round. Statistics Netherlands now conducts for the fourth time a virtual census. However, the Dutch data that have been compiled on 1981 and 1991 were of a much more limited character than the set of tables of the 2001 and 2011 censuses. Moreover, they were largely based on a register count of the population in combination with the then existing surveys about the labour force and housing conditions.

31. The technique of repeated weighting has been used successfully to produce a consistent set of tables for the 2001 census and will also be the key method for the 2011 census. Every table is calculated from the largest block from which the table can be determined. All tables are estimated this way for the census with the correct weights, and therefore the tables' results are mutually consistent. By starting every time from the largest block, the most detailed possible census tables have been achieved. Before compiling tables with this technique, micro-integration of the different sources in the SSD remains important. In the micro-integration process the data are checked and incorrect data are adapted. It is strongly believed that micro-integrated data will provide more reliable results, because they are based on a maximum amount of information. Also the coverage of subpopulations will be better, because when data are missing in one source, another source can be used. Another advantage of micro-integration and repeated weighting is that there is no reason for confusion among users of statistical information anymore, because there will be one figure on each socio-economic phenomenon, instead of several figures depending on which sources have been used.

References

Corbey, P., 1994. Exit the population Census. *Netherlands Official Statistics*, Volume 9, summer 1994, pp. 41-44.

Duin, C. van and V. Snijders, 2003. Simulation studies of repeated weighting. Discussion paper 03008, Statistics Netherlands, Voorburg / Heerlen. http://www.cbs.nl/NR/rdonlyres/203C85C6-7075-47A0-97BA-A3B748D393FE/0/Discussionpaper03008.pdf

European Commission, 2008. Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses. Official Journal of the European Union, L218, pp. 14-20.

European Commission, 2009. Commission Regulation (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdown. Official Journal of the European Union, L329, pp. 29-68.

European Commission, 2010a. Commission Regulation (EU) No 519/2010 of 16 June 2010 adopting the programme of the statistical data and of the metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council. Official Journal of the European Union, L151, pp. 1-13.

European Commission, 2010b. Commission Regulation (EU) No 1151/2010 of 8 December 2010 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission. Official Journal of the European Union, L324, pp. 1-12.

Houbiers, M., 2004. Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics*, Volume 20, No. 1, pp. 55-75.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders, 2003. Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Statistics Netherlands, Voorburg / Heerlen. http://www.cbs.nl/NR/rdonlyres/6C31D31C-831F-41E5-8A94-7F321297ADB8/0/discussionpaper03005.pdf

Knottnerus, P. and C. van Duin, 2006. Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, Volume 22, No. 3, pp. 565-584.

Kroese, A.H. and R. H. Renssen, 2000. New applications of old weighting techniques, constructing a consistent set of estimates based on data from different sources. *ICES II, Proceedings of the second international conference on establishment surveys, survey methods for businesses, farms, and institutions, invited papers*, June 17-21, 2000, Buffalo, New York, American Statistical Association, Alexandria, Virginia, United States, pp. 831-840.

Mushkudiani, N., J. Daalmans and J. Pannekoek, 2012. Macro-integration techniques with applications to census tables and labour market statistics. Discussion paper, Statistics Netherlands, The Hague / Heerlen. http://www.cbs.nl/NR/rdonlyres/AD653253-647D-4FFD-AFC4-67E2BDE602EE/0/201201x10pub.pdf

Rao, J.N.K., 2003. Small area estimation. Wiley, New York, United States.