Assessing the quality of register-based population censuses: processes and outputs

Bart F.M. Bakker[1]
Statistics Netherlands, The Hague, Netherlands – bfm.bakker@cbs.nl
VU University, Amsterdam, Netherlands

Abstract

Recently emerged alternatives to the field enumeration approach for censuses are pure register-based or mixed register-survey approaches. Key issues that have to be resolved: under and over coverage, measurement error and linkage error. Under coverage can be determined and corrected for with capture-recapture methods. Over coverage can be determined and corrected for by removing records of people that do not belong to the population and duplicates by linking the records in a (quasi) population register to each other. Measurement error can be estimated by Structural Equation Models (for numerical variables) and Latent Class Analysis (for categorical variables) with a measurement component if another source is linked to the census data that measures the same concept. Linkage error can be estimated using probabilistic linkage methods. However, none of these methods is error free in itself.

Keywords: capture-recapture methodology, latent class analysis, linkage error, census quality

1. Introduction

The importance of census outcomes can hardly be overstated. Census information is used to substantiate government policies as it gives a very detailed picture of society and its social and regional differences. Moreover, census outcomes are important sources for historical trends longer than a few decades. Finally, because of their relatively large consistency between countries, census data are increasingly used for international comparative studies. The success of the Integrated Public Use Microdata Series (IPUMS) proves that this development is substantial. IPUMS consists of 277 micro data samples from census records from 82 countries from all around the world (Minnesota Population Center 2017). Therefore, the quality of census outcomes should be high. That triggers the question what the quality is of censuses.

In most countries, a census is organised by interviewing from door to door. This is the traditional way of census taking. Since the census of 1980 an increasing number of countries used administrative data to collect the necessary information. The most important reason for this is that it is much cheaper. Moreover, it usually takes a long time to process the data collected with a traditional census while the processing time of a register-based census is much shorter, in particular if the registers are already used for the regular production of official statistics.

There are two models for the register-based census. The first one is that it uses only register information, the second one is that register data are combined with survey data to add the variables that are not available in registers. Denmark was the first country in the world with a completely register-based census as early as 1980. In 1990, Finland was the next one and thus reduced the costs for the census by more than 90% between 1980 and 1990 (Ruotsalainen 2011). The 2011 census is exclusively register-based in the Nordic countries, Austria, Belgium, Slovenia, and Switzerland, while Germany, Netherlands, Latvia, Lithuania, and Israel rely heavily on registers (UNECE 2014; Bechtold

[1] Bart F.M. Bakker is head of the Methodology Department of Statistics Netherlands in The Hague and professor in register methodology at VU University Amsterdam. The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author thanks Jan van der Laan and Sander Scholtus for their critical review of this paper.

2013). The costs for register-based censuses are much lower than the costs of traditional censuses: for example, the 2011 census in Denmark cost only $0.07 per head of the population, compared to $40.17 for the US Census (UNECE 2014, 64).

A framework for the errors in register-based statistics has been developed by Bakker and Daas (2012) and Zhang (2012). As in the total survey error approach (Groves et al. 2007), they have defined possible error sources for each step in the process of register-based statistics. The process of a register-based census typically is that a population register or a combination of registers are used to produce a list of the population on census date. We call the result of this step (quasi) population register (qPR). After that, the other necessary registers and surveys are linked, the data are edited, the target variables are deduced and a solution is applied to the missing values in these variables (e.g. introducing a category unknown or imputation). The last step is the estimation of the outcomes. In the case of a fully register-based census, this is merely simple counting. However, if a combination of registers and surveys is used, some estimation procedures has to be applied (e.g. weighting, consistent repeated weighting, macro-integration techniques). All steps contribute to the total error in the census outcomes.

Bakker and Daas (2012) and Zhang (2012) distinguish between representation and measurement error. A dataset contains representation error if the data do not represent the target population. It contains measurement error if a variable measures something else than the target variable (or has a large amount of uncertainty, but that is in the case of a register-based census seldom a problem). An inevitable step in the process of a register-based census is that different registers, and sometimes surveys are linked. In this paper, I discuss the different sources of representation and measurement error based on this framework, the methods to estimate the size of these errors and, if available, the methods to adjust the outcomes for these errors.

2. Representation error

In most countries, in the first step, the target population is deduced from a population register or different linked registers. The representation error in such a qPR consists of three kinds of error: people are in the qPR while they do not belong to the target population, people are missing from the qPR and there are duplicates in the qPR who do belong to the target population. The target population of the 2011 round in the EU was the usual resident population. According to the (United Nations Statistics Divisions, 2008) we can define usual residence as:

> *"1.461. In general, "usual residence" is defined for census purposes as the place at which the person lives at the times of the census, and has been there for some time or intends to stay there for some time"*

According to the European Parliament (2013), Regulation (EU) No 1260/2013 of the European Parliament, usual residence is defined as:

> *"The place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage"*

"Some time" has been defined as one year.

Records from people who are in the (quasi) population register while they do not belong to the population have to be removed. Of course, this depends on the possibilities of identifying those people and determining the date that they became a resident. In register data, it is mostly possible to deduce the starting date that people are registered as living, working or following education in the country. However, identification and determining the starting date will certainly be not perfect because in practice the data quality of the persons who do not belong to the population is not as high as we want it to be, often containing a relatively large number of missing values.

The second error is that people are missing in the (quasi) population register. The missed number of usual residents can be estimated by means of capture-recapture methods (Bishop, Fienberg and Holland, 1975; International Working Group for Disease Monitoring and Forecasting. 1995; Van der Heijden et al., 2012; Gerritse et al., 2015). This method can also be used to estimate the under coverage of traditional censuses. The method can be applied if besides the (quasi) population register a second source is available (be it a register or a survey), link this second source to the (quasi) population register and apply a log-linear model under the assumption that the two sources are independent. To get accurate outcomes from these models, several assumptions have to be met. In practice these assumptions cannot always be met, particularly when data are originally not collected for statistical purposes. Violation of the assumptions leads to severely biased results (Brown et al., 2006; Gerritse, 2016)

However, there are a few guidelines to diminish the impact of violation of the assumptions. The first assumption that has to be met is that inclusion of persons in the first data source is independent of inclusion in the second source. This assumption can be relaxed in two ways: by linking a third source or by adding covariates to the model. If both guidelines are followed, the assumption is relaxed to the assumption that the higher order interaction of the three inclusion probabilities is zero within each cell of the covariates. The second assumption is that the population is closed during the data collection period. This assumption can be easily met by using data of only one reference date. If that is not possible, then take data that is collected in a very short period. The third assumption is that the sources are perfectly linked. False negative and false positive links lead to biased estimates. Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015) have developed methods to adjust the outcomes of capture-recapture analysis for linkage error. An important condition to apply this method is that you use probabilistic linkage to link your sources. It is based on the idea that you make use of the estimated probabilities that linked pairs are correct links. These improvements of the capture-recapture method seem promising.

One should verify the possibility of duplicate records in the qPR. Duplicate records can be detected on the basis of identifying variables. The amount of success of this step depends on the quality of the identifying variables. If a lot of missing values appear in the identifying variables, we expect that the remaining records contain to some extent duplicates. If all individuals are identified by a personal identification number correctly, this should not lead to large quality problems. However, if a personal identification number is missing or is not complete, address information is necessary to identify individuals. When registers of different quality and in particular different levels of administrative delay are combined, individuals who move from one place to the other can be seen as multiple different persons. A possible solution is to use additional information on the history of removals. If that is not available, only expert guesses based on knowledge of the quality of the different sources are possible to get an idea how many duplicates remain. If that is substantial, then it would be advisable to give those records that are suspected to be a duplicate a weight of one half.

3. Measurement error

A register-based census contains measurement error if the register variables are conceptually different from the target variables, or if they are measured with a systematic or random measurement error. Because most of the register information is collected with the use of traditional survey techniques, it is naïve to assume that registers do not contain systematic or random measurement error. To determine the size of the measurement error, two methods are available for different levels of measurement. The first method is a Structural Equation Model (SEM) with a measurement component for continuous

variables and the second method is Latent Class Analysis (LCA) for categorical variables. The basic idea is that you use (at least) two measures of the same concept and conceive the association between each observed variable and the target variable as a measure of the quality of that observed variable . This association is known as the indicator validity of the observed variable. This can be done by linking a (small) survey collected by interviewing to the combined registers or by linking another register that contains the same variable(s).

If one uses a SEM to estimate the size of the measurement error, one has to distinguish between the error in the estimated relationships between the variables, the indicator validity, and the intercept bias, i.e. the mean of the variables. An example of the application of a SEM for the indicator validity is Bakker (2012). He linked a survey to a combination of a large number of registers in order to determine the quality of main variables as age, gender, education attainment and wages. By applying the classical test theory, the validity can be determined by using linked survey and administrative data which should measure the same concepts. A linear structural equations model with a measurement component is used to compute the indicator validity. The analyses reveal that age and gender are almost measured perfectly in both sources, educational attainment is better measured in the register data and wages in the survey data. Scholtus, Bakker and Van Delden (2015) introduce an additional test on the intercept bias, using similar models. However, in order to pin the “true level” a small audit sample is needed of flawless quality. Using the indicator validity and the intercept bias, it is possible to adjust the outcomes of the census data.

However, SEM can only be used if one has continuous conceptual variables. For instance educational attainment can be considered a continuous concept, but it can only be measured as a categorical variable (in years, or levels). In censuses most of the variables are categorical. In that case errors are classification errors: an individual is wrongly classified to a category. For the determination of these classification errors one can apply LCA. An example of this approach is Pavlopoulos and Vermunt (2015) with an application to estimating permanent and temporary employment in the Netherlands. They use linked longitudinal categorical data from a register and the Labour Force Survey and apply a particular group of latent class models called Hidden Markov Models (HMMs). HMMs are applied to describe a turnover or transition in some characteristic assuming that it is driven by a process without memory and that it is measured with an error. These models probabilistically estimate the latent states at the individual level and provide also estimates for the distribution of these states as well as for the mobility between them. They show that the latent transition rate from temporary to permanent employment in the Netherlands is less than half than our observed data suggest. LCA leads to a classification table which can be used to correct for the classification error.

4. Linkage error

After the target population has been determined the other registers are linked to the dataset. In the last step, for those countries who use survey data these data are linked to the census dataset. Linkage error could then lead to both representation error (duplicate records, false negative links) and measurement error (false positive links). Moreover, the combination of register and survey data can lead to inconsistent outcomes.

False negatives exist if records that belong to the same person (or dwelling) were not linked. If a census is based on a qPR, this means that variables in sources that are linked to this register are missing. This can lead to severe quality problems. E.g. if an administrative source with information on jobs is linked to the qPR, false negatives lead easily to an underestimation of the number of individuals belonging to the workforce. It is very difficult to determine whether a job should be linked or not because the job register describes the entire population and also includes individuals not belonging to the population (e.g. cross border workers, temporary workers). If it is possible to identify

the cross border and temporary workers and those who should have been linked, one can try to correct for the false negatives by weighting.

False positives exist if records that belong to different persons (or dwellings) were linked. This leads to a specific form of measurement error. The linkage process should focus on the minimisation of the false negatives and false positives. However, if a large number of false negatives remain after a simple deterministic linkage process in which all the identifying variables are identical, most countries apply a form of probabilistic linkage in which some deviation from total similarity is allowed. That leads inevitably also to some false positives. If the resulting file can be linked to a survey, one can apply LCA or SEM to estimate the quality of the estimations and to correct for measurement error as is described in section 3.

5. Conclusions

Register-based censuses have become more and more popular, mainly because of the low costs. However, census outcomes are very important for decision making and the quality should therefore be high. There are three interdependent errors in register-based censuses: under and over coverage, measurement error and linkage error. Under coverage can be determined with capture-recapture methods. Over coverage can be estimated by linking the records in a (quasi) population register to each other. Measurement error can be estimated by Structural Equation Models (for numerical variables) and Latent Class Analysis (for categorical variables) with a measurement component if another source is linked to the census data that measures the same concept. Linkage error can be estimated using probabilistic linkage methods. For each error source, it is possible to make some adjustments to the data. However, because the error sources are interdependent, it is to be seen whether it is possible to design a process that leads to the optimal outcomes of the register-based census. Moreover, more work has to be done on estimating accuracy and reliability measures of these outcomes. These are few of the challenges that are still ahead.

Another aspect that has not been mentioned yet is the solution to the problem of inconsistent estimations that could be caused by the combination of register and two or more sample surveys. This is a field of research that has caught much attention lately. However, this paper is restricted to representation and measurement error in particular the errors that lead to biased outcomes. Diminishing inconsistencies is the last step in the estimation process and should not lead to substantially different estimates.

References

Bakker, B.F.M., 2012. “Estimating the validity of administrative variables.” Statistica Neerlandica, 66: 8-17.

Bakker, B.F.M. and P.J.H. Daas. 2012. “Methodological Challenges of Register-Based Research.” Statistica Neerlandica 66: 2–7.

Bechtold, S. 2013. “The New Register-Based Census of Germany – a Multiple Source Mixed Mode Approach.” In Proceedings of the World Statistics Congress, August 25-30, 2013, Hong Kong (pp. 259–264). Available at: http://2013.isiproceedings.org/Files/IPS027-P2-S.pdf (last accessed June 19, 2015)

Bishop, Y., S. Fienberg, and P. Holland. 1975. Discrete Multivariate Analysis, Theory and Practice. New York: McGraw-Hill.

Brown, J., O. Abbott, and I. Diamond. 2006. “Dependence in the 2011 One-Number Census Project.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 169: 883–902.

Di Consiglio, L. and Tuoto, T., 2015. “Coverage Evaluation on Probabilistically Linked Data”, Journal of Official Statistics, 31: 415–429.

Ding, Y. and S.E. Fienberg. 1994. “Dual System Estimation of Census Undercount in the Presence of Matching Error.” Survey Methodology 20: 149–158.

European Parliament 2013. Regulation (ec) no 1260/2013 of the European parliament and of the council of 20 November 2013 on European demographic statistics. Official Journal of the European Union L 330/39

Gerritse, S.C., 2016. An application of population size estimation to official statistics. Sensitivity of model assumptions and the effect of implied coverage (Ph. Dissertation Utrecht University).

Gerritse, S.C., P.G.M. van der Heijden and B.F.M. Bakker, 2015. “Population size estimation for violating parametric assumptions in loglinear models”, Journal of Official Statistics, 31: 357-379.

Groves, R. M., F. J. Fowler JR., M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, 2007. Survey methodology (Wiley Interscience, New York)

International Working Group for Disease Monitoring and Forecasting. 1995. “Capture-Recapture and Multiple Record Systems Estimation. Part I. History and Theoretical Development.” American Journal of Epidemiology 142: 1059–1068.

Minnesota Population Center. 2017. Integrated Public Use Microdata Series, International: Version 6.2 [Machine-readable database]. Minneapolis: University of Minnesota. Available at: https://international.ipums.org/international/ (last accessed 3 March 2017)

Pavlopoulos, D., & Vermunt, J. K. 2015. Measuring temporary employment. Do survey or register data tell the truth? Survey Methodology, 41: 197–214.

Ruotsalainen, K. 2011. A census of the World Population is Taken Every Ten Years (Helsinki: Statistics Finland). Available at: http://tilastokeskus.fi/tup/vl2010/art_2011-05-17_001_en.html (last accessed 3 March 2017).

Scholtus, S., Bakker, B.F.M. & A. van Delden, 2015, *Modelling measurement error to estimate bias in administrative and survey variables. Discussion Paper 2015-17* (Den Haag / Heerlen: Statistics Netherlands)

United Nations Statistics Division (2008). Principles and Recommendations for Population and Housing Censuses - Rev.2. Statistical papers Series M No 67/Rev.2. United Nations, New York.

UNECE (United Nations Economic Commission for Europe). 2014. Practices of UNECE Countries in the 2010 Round of Censuses. New York: United Nations.

Van der Heijden, P.G.M., J. Whittaker, M.J.L.F. Cruyff, B.F.M. Bakker, and H.N. van der Vliet. 2012. “People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Estimates and the Role of Active and Passive Covariates.” The Annals of Applied Statistics 6: 831–852.

Zhang, L.-C. 2012. “Topics of Statistical Theory for Register-Based Statistics and Data Integration.” Statistica Neerlandica 66: 41–63.