DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 349–355, http://dx.doi.org/10.1515/JOS-2015-0021

# Preface

## 1. Introduction to the Special Issue on Coverage Problems in Administrative Sources

Administrative data are being used more and more in official statistics and academic research as an alternative to interviewing, in particular for census taking. An important issue with the use of administrative sources for statistical purposes is that they often suffer from under- and overcoverage with respect to the population of interest. The articles in this special issue focus on methodologies for dealing with these coverage problems. A common theme in many of the articles is that they address the assumptions behind the dual system capture-recapture methodology that is often used to correct for undercoverage in censuses – either by evaluating the robustness of this method to violations of certain assumptions or by proposing new methods that relax some of these assumptions.

## 2. The Importance of Administrative Data

In many countries the use of administrative data has been stimulated by the fact that census information is vital and at the same time very expensive if the data are collected by door-to-door interviewing.

The importance of a census 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 238 microdata samples from census records from 74 countries from all around the world (Minnesota Population Center 2013).

However, census taking by door-to-door interviewing is very costly. In the United States (US), the cost of the 1990 Census was $2.6 billion and this increased to $13 billion in 2010. The costs of conducting a US census have more than doubled every ten years.

**Acknowledgments:** The idea of this special issue on coverage arose at a meeting in 2012 in Örebro, Sweden, where specialists in the field of official statistics discussed quality issues of administrative data. It used the papers presented at the 59th World Statistics Congress in Hong Kong in 2013 as an important source. We want to thank the contributors, reviewers and discussants for their efforts to make this a successful issue of JOS. We also want to thank the editorial staff at JOS for giving us the opportunity to compile this special issue and for their support and guidance along the way. We hope this issue provides interesting reading.

© Statistics Sweden

In England and Wales, the door-to-door census of 2011 cost was 482 million British pounds. The 2001 census cost was less than half that amount: 210 million pounds (*Economist*, 2 June 2011).

Therefore, countries are looking for more cost-effective alternatives. One popular way to reduce costs is to make use of administrative records like population, tax, or health registers, and, if these sources do not cover all information that is needed, to combine these sources with data from sample surveys. Denmark was the first country in the world to conduct a completely register-based census as early as 1980. In 1990, Finland was the next to follow 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 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).

## 3. Coverage Problems Defined

Censuses are very important for giving a detailed picture of the social and regional differences in each country. To fulfil that role, they should cover the entire population and only the population. However, both a traditional census and a register-based census have coverage problems. The traditional census could miss parts of the population due to incomplete address files and nonresponse. Register-based censuses could miss parts of the population because not all elements of the population are registered. In both cases, this might lead to undercoverage. Another problem is that registers erroneously include individuals that are no longer part of the population. This leads to overcoverage. This could be the case, for example, if removals, emigrations and deaths have been registered with a certain time lag. Administrative delay is an important source of error in administrative data (Bakker and Daas 2012; Zhang 2012).

The usual way of census coverage evaluation is to conduct a postenumeration survey (or coverage measurement survey) to the census data in order to estimate the total population size using capture-recapture methods. For that purpose, a register could also be used instead of the postenumeration survey. This is also known as dual-system estimation (e.g., Hogan 1993; Brown et al. 2006; Chen et al. 2010; Sadinle and Fienberg 2013; Baffour et al. 2013). In most cases, log-linear models are used to estimate the size of the population and the part missed by the observed data.

The quality of the outcomes of capture-recapture methods with two sources rely on five assumptions (Bishop et al. 1975; International Working Group for Disease Monitoring and Forecasting 1995):

1. The probability of being in the second source does not depend on the probability of being in the first source.
2. The probabilities are homogeneous across all elements in at least one source, or, if probabilities are heterogeneous in both sources, the sources of heterogeneity are unrelated (see Van der Heijden et al. 2012).

3. The population is closed, that is, there are no individuals entering or leaving the population during the period of observation.
4. The elements of the population in the two sources can be perfectly linked.
5. There are no erroneous captures in either the first or second source.

Violating these assumptions can cause severe bias in the population size estimates. In particular, violation of perfect linkage and independence can lead to serious bias (Brown et al. 2006; Baffour et al. 2013; Sadinle and Fienberg 2013).

To fulfil the needs of the main users of the census, the information on the total population should have all the details, much more detail than the cross table of the covariates. These needs can be fulfilled by weighting the data of individuals in the census, be it a traditional door-to-door census or a register-based census. Here the estimation of the total population by the cross table of the covariates in the log-linear model can be used as a weighting frame for the construction of the weights. The success of this procedure depends on the association between the variables used for the construction of the weights and the target variables on the one hand, and the probability of being missed in the administrative data on the other hand, because it is similar to weighting procedures correcting for selective nonresponse in household surveys. The higher the associations, the better the estimates become (Särndal et al. 1992, 588-589; Bethlehem et al. 2011, 207-246).

An increasing number of countries use administrative data not only for census purposes, but also for their regular production of official statistics and for academic research. The coverage problems that occur in the register-based censuses are similar to other fields of interest. In this special issue, we present a number of methodological studies that address important aspects of the methodological problems in estimating population sizes and other official statistics with administrative data and suggest solutions for some of them.

## 4. In this Issue

Nine studies are presented, each dealing with specific aspects of the methods for estimating under- or overcoverage. All studies deal with undercoverage, and several deal with overcoverage as well.

Gerritse, Van der Heijden, and Bakker study undercoverage of linked data sources and methods to remedy this using dual-system estimation. The sensitivity of the population size estimates is studied for violation of the assumption that in dual-system estimation the inclusion probabilities of two sources are independent (this is Assumption 1 discussed above). They simulated this with or without covariates, using log-linear models with offsets. In their simulation with real data they found that under certain circumstances, this sensitivity is high and leads to implausible results. If the first source has a better coverage than the second source, then the sensitivity is higher compared to when the coverage of the first source is lower. They also studied models in which a covariate is only available in one of the two sources, which is a rather common situation. They show that, in accordance with Zwane and Van der Heijden (2007) and Van der Heijden et al. (2012), ignoring covariates that are related to the inclusion probability may lead to biased estimates.

If overcoverage occurs, there are erroneous captures in either the first or second source or in both sources. This is a violation of Assumption 5 of dual-system estimation discussed above. The article of Zhang proposes models that take into account both over- and undercoverage. His models are developed for (i) two lists that may both have over- and undercoverage and (ii) an additional coverage survey. Assumptions are that the additional coverage survey has only undercoverage, and that the additional coverage survey can be completely linked to the two lists. Simulations suggest the usefulness of the models proposed and this may prove to be a promising direction for solving applied problems where overcoverage plays a role. The models also deal in some way with Assumption 3 discussed above, that of a closed population.

When administrative data are used for the census or other official statistics, most of the time different administrative sources are combined to produce the desired tables. However, record linkage is not an error-free process. Missed links can lead to undercoverage and incorrect links can lead to overcoverage (Bakker and Daas 2012). Both missed links and incorrect links are violations of the abovementioned Assumption 4, which states that individual records can be perfectly linked. There has been an explosion of record-linkage applications, yet there has been little work on making correct inference using such linked files. When the possible existence of these errors is not taken into account, however, this may lead to biased inferences. Chipperfield and Chambers develop a method of making inferences for the measurement of binary variables in the population when record linkage is not an error-free process. In particular, they develop a parametric bootstrap approach to estimation which can accommodate sophisticated probabilistic record linkage techniques that are widely used in practice (e.g. 1-1 linkage, i.e., where every record on one file is linked to a distinct and different record on the other). The article demonstrates the effectiveness of this method with a simulation and an application to real data.

Another article on linkage, and hence on a violation of Assumption 4, is provided by Di Consiglio and Tuoto. They build on earlier work by Ding and Fienberg (1994). Ding and Fienberg proposed estimators corrected for linkage bias, and Di Consiglio and Tuoto provide a generalization of these estimators. The method is illustrated with an application to real data to estimate the number of casualties due to road accidents, integrating data from two registers. Simulated data are used to show the benefit of the proposed new method over the existing estimators.

In England and Wales, several alternatives to a traditional census have been evaluated in the Beyond 2011 programme. The recommended option for 2021 makes use of administrative data. In England and Wales, the National Health Service Patient Register (NHSPR) is the most comprehensive administrative source. It covers everyone registered with a general practitioner (GP). However, it is known that direct estimates from the NHSPR of the population size by sex, age, and region are biased due to a variety of problems, such as administrative delays when people change GPs, persons being registered more than once, and so on. This may be seen as ‘local’ overcoverage and hence as a violation of Assumption 5. Yildiz and Smith determine which population groups are not well presented in the NHSPR and propose a method for correcting for the inaccuracies. For this purpose, they combine the NHSPR with marginal information on sex, age, and region from an auxiliary source, which is supposed to provide unbiased estimates at a

regional level, keeping the higher-level interaction structure intact. Population counts are estimated by using different log-linear models with offsets that take care of the interaction structure. In their application, they use auxiliary information from the 2011 Census. However, in the future marginal information from other data structures may be used to correct for bias in the NHSPR.

In the study of Blackwell, Charlesworth, and Rogers, the quality assurance of the 2011 Census of England and Wales is discussed. This quality in terms of coverage has been determined by linking the traditional census data to administrative sources. The Office for National Statistics (ONS) has invested a lot of effort in the process of linking those data. The linking strategy reflected the hierarchical structure of people living within and across addresses and included evidence from the census field operation. Patterns of differential coverage in the different administrative sources emerged.

Bryant and Graham have a different approach to deriving population estimates from multiple administrative data sources with undercoverage. They do not combine different administrative sources at the individual level by record linkage, but at an aggregate level: the cell count. The overall model contains submodels describing regularities within demographic processes and the relations between the demographic processes and the available datasets. They use Bayesian methods, because this makes it possible to account for different sources of uncertainty. Coverage rates are used as a diagnostic and as an important source to weight the data. They apply this method to data from New Zealand and try to estimate the population by age (5-year groups), sex, time and region. The process of deriving the weights is automatic and data driven. They show that their approach is promising, in particular if for some reason you are not able to perform high-quality record linkage at the individual level.

A final article deals with coverage but is not directly linked to the population census. Coverage problems could also occur if units are wrongly classified, for example if addresses are wrongly classified by region. This can lead to a net undercoverage if the balance between erroneously assigned units and erroneously unassigned units is negative, and it can lead to a net overcoverage if this balance is positive. The study of Burger, Van Delden, and Scholtus applies a resampling method to assess the sensitivity for source-specific classification errors in mixed-source statistics, such as an enterprise register and survey. The method can be used for deciding how to allocate resources in the production process of statistics. They applied the method to short-term business statistics suggesting that shifting classification resources from small and medium-sized enterprises to large ones may have no effect on the accuracy, because the gain in precision is offset by the creation of bias.

At the end of this issue, Raymond Chambers, Anders Holmberg, and Stephen Fienberg tie these manuscripts together in insightful ways. Chambers focuses on the articles of Burger et al., Gerritse et al., Di Consiglio and Tuoto, and Zhang, which have in common that they deal with measurement error methodology for official statistics. He argues that the difference is that Burger et al. and Gerritse et al. only point out deficiencies when assumptions are not met, but that Di Consiglio and Tuoto and Zhang try to come up with solutions. Anders Holmberg comments on all articles from the perspective of the tasks of offices of national statistics and his own personal experiences. Fienberg discusses all contributions and provides additional links of these articles to the literature, to work on

official statistics in the U.S. as well as to his own work. Moreover, he sketches a research programme to continue the research on the most important topics discussed in this special issue. It is definitely worth taking the time to study these comments in addition to the contributions of the authors.

*Bart F.M. Bakker*
Team Methodology, Statistics Netherlands,
and VU University, The Netherlands.
Email: bfm.bakker@cbs.nl

*Peter G.M. van der Heijden*
Department of Methodology and Statistics,
Utrecht University, The Netherlands,
and University of Southampton, UK.
Email: P.G.M.vanderHeijden@uu.nl

*Sander Scholtus*
Team Methodology, Statistics Netherlands,
The Netherlands.
Email: s.scholtus@cbs.nl

## 5. References

Baffour, B., J. Brown, and P.W.F. Smith. 2013. "An Investigation of Triple System Estimators in Censuses." *Statistical Journal of the IAOS* 29: 53–68.

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).

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley & Sons.

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.

Chen, S.X., C.Y. Tang, and V.T. Mule, Jr. 2010. "Local Post-Stratification in Dual System Accuracy and Coverage Evaluation for the US Census." *Journal of the American Statistical Association* 105: 105–119.

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

Economist 2011, "Old Style Censuses are Cumbersome and Costly. Reform is coming." 2011. *The Economist*, June 2. Available at: http://www.economist.com/node/18772674 (last accessed 19 June 2015) .

Hogan, H. 1993. "The Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060.

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. 2013. *Integrated Public Use Microdata Series, International: Version 6.2* [Machine-readable database]. Minneapolis: University of Minnesota.

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 11 September 2013).

Sadinle, M. and S.E. Fienberg. 2013. "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." *Journal of the American Statistical Association* 108: 385–397.

Särndal, C.-E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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.

Zwane, E.N. and P.G.M. van der Heijden. 2007. "Analysing Capture-Recapture Data when Some Variables of Heterogeneous Catchability are not Collected or Asked in All Registries." *Statistics in Medicine* 26: 1069–1089.