JOS
Journal of Official Statistics

September 2013

Journal of Official Statistics
vol. 31, i. 3 (2015)

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.

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.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 357–379, http://dx.doi.org/10.1515/JOS-2015-0022

Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models

Susanna C. Gerritse[1], *Peter G.M. van der Heijden*[2], *and Bart F.M. Bakker*[3]

An important quality aspect of censuses is the degree of coverage of the population. When administrative registers are available undercoverage can be estimated via capture-recapture methodology. The standard approach uses the log-linear model that relies on the assumption that being in the first register is independent of being in the second register. In models using covariates, this assumption of independence is relaxed into independence conditional on covariates. In this article we describe, in a general setting, how sensitivity analyses can be carried out to assess the robustness of the population size estimate. We make use of log-linear Poisson regression using an offset, to simulate departure from the model. This approach can be extended to the case where we have covariates observed in both registers, and to a model with covariates observed in only one register. The robustness of the population size estimate is a function of implied coverage: as implied coverage is low the robustness is low. We conclude that it is important for researchers to investigate and report the estimated robustness of their population size estimate for quality reasons. Extensions are made to log-linear modeling in case of more than two registers and the multiplier method.

Key words: Capture-Recapture methodology; dual-system estimation; sensitivity analysis; census; Poisson log-linear regression.

1. Introduction

For the Census of 2011, an increasing number of countries used administrative data to collect the necessary information. Under census regulations a quality report is obligatory, and one of the aspects that needs to be addressed is the undercoverage of the census data. This asks for an estimate of the size of the population. If one wants to estimate the size of a population, capture-recapture methods, making use of log-linear models, are commonly used (Fienberg 1972; Bishop et al. 1975; Cormack 1989; International Working Group for Disease Monitoring and Forecasting 1995). These methods go by different names, such as mark-recapture methods, dual-system methods or dual-record system methods. In this

[1] Utrecht University, Methods and Statistics, Padualaan l4, Utrecht 3584 CH, The Netherlands and University of Southampton, UK. Email: sc.gerritse@gmail.com
[2] Utrecht University, Methods and Statistics, Padualaan l4, Utrecht 3584 CH, The Netherlands and University of Southampton, UK. Email: P.G.M.vanderheijden@uu.nl
[3] Statistics Netherlands, Methodology, P.O.Box 24500, 2490 HA, The Hague, The Netherlands and VU University, Netherlands. Email: bfm.bakker@cbs.nl

Acknowledgments: An earlier version of this article was presented at the 59th World Statistics Congress, 25 – 30 August 2013. Rik van der Vliet and Peter-Paul de Wolf for their valuable comments on earlier drafts of this article. Olav Laudy for his help in writing the SPSS routine and Raymond Chambers for pointing out the link to the work of James Brown. We would also like to express our gratitude for the valuable comments of our reviewers and editors.

article we use the label capture-recapture. In countries with a traditional census a postenumeration survey could be organised to collect recaptured data, as was the case for instance in the United Kingdom (Brown et al. 1999; ONS 2012), and in the U.S. (Wolter 1986; Bell 1993; Nirel and Glickman 2009). In this case, a survey with a relatively small sample size is linked to the census data. In countries with a census based on administrative data, the approach used most is to find two registers and treating these as the captured and recaptured data. The method includes linking the individuals in the registers and subsequently estimating the number of individuals missed by both registers.

However, the outcome of the capture-recapture method depends heavily on some assumptions underlying the data. In particular, if two sources are used, it is usually assumed that inclusion in the captured data is independent of inclusion in the recaptured data. A second assumption deals with homogeneity versus heterogeneity of inclusion probabilities. If there is one source of heterogeneity it is assumed that at least for one of the two sources the inclusion probabilities are homogeneous (Chao et al. 2001; Zwane and Van der Heijden 2004). If there are two sources of heterogeneity (two covariates), the estimates are unbiased if the inclusion probabilities of the first source vary with one source of heterogeneity, and the inclusion probabilities of the second source vary with a second source of heterogeneity, but the two sources of heterogeneity are statistically independent (Seber 1982, 86). The remaining two assumptions are that the population is closed and that the registers are perfectly linked.

The assumption of independence between two registers is very strict and can easily be violated. Under dependence between registers, the inclusion probability of one register is related to the inclusion probability of the other register. Then, under positive dependence individuals in the captured data have a higher probability of also being in the recaptured data, resulting in an underestimation of the population size estimate. Additionally, under negative dependence the opposite holds (Hook and Regal 1995).

Independence is an unverifiable assumption, that is, it cannot be verified from the data used for the estimation of the population size. The log-linear independence model for the linked captured and recaptured data has three parameters, whereas there are only three counts. Because the observed counts are equal to fitted counts, the independence model is the saturated model (compare van der Heijden et al. 2012). Thus we cannot assess dependence from the saturated model. One way of reducing the impact of the strict independence assumption is to replace it with the lesser strict assumption of independence conditional on covariates. Adding covariates enables us to reduce heterogeneity introduced to the model due to the specific covariate, adjusting the population size estimate for the better. The situation of a saturated model also holds when covariates of individuals are taken into account and we operate under the log-linear conditional independence model. However, we are interested in what the impact of mild or severe violations of (conditional) independence is on the population size estimate. It does not necessarily have to be the case that violation of the (conditional) independence assumption results in a substantive bias in the population size estimate. It is of important to also assess what happens when the other assumptions are violated. However, looking at all assumptions at once is very complex. In this article, we will thus focus on the violation of the independence assumption, assuming all other assumptions to be met.

We propose a general approach to sensitivity analyses under the log-linear model framework using a log-linear Poisson regression, a special case of the generalized linear model. Where in the saturated model specific interaction parameters are equal to zero, we impute fixed values departing from zero for these parameters, thus simulating dependence, and investigate the impact on the population size estimate. As the log-linear interaction parameters are closely related to the (conditional) odds ratio, there is a clear interpretation for the values to which we fix the parameters.

Similar findings come from the research of Brown et al. (1999), where the census was linked to a Post Enumeration Survey to assess under- and overcoverage (cf. also Wolter 1986; Bell 1993). Brown et al. (1999) used a fixed odds ratio of 0.1 and 10 to investigate the impact of simulated dependence on the population size estimate. They showed that fixed dependence can seriously bias the population size estimate under the independence assumption. Results like these are valuable, since they give insight to the size of the impact of violated independence. However, research into the robustness of the population size estimator under violation of independence is non standard. As far as we know, other research on the impact of the violation of independence involves simulation studies, an already known population size estimate or uses multiple sources (Wolter 1986; Bell 1993; Cormack et al. 2000; Hook and Regal, 1992, 1997, 2000; Brown et al. 2006; Baffour et al. 2013).

We extend the results of Brown et al. (1999) by, instead of using the standard log-linear model, working under a log-linear Poisson regression where we simulate a fixed dependence using offsets. In simulating dependence by adding a fixed offset value to the log-linear model, we can compare the population size estimate under independence to the population size estimate under a ‘true’ dependence. Additionally we extend our two-register independence model to the case with covariates observed in both registers (fully observed covariates) and covariates observed in only one register (partially observed covariates).

Partially observed covariates are usually ignored because including them would lead to missing values in the other register. However, ignoring these covariates when they actually are related to the inclusion probability of the register results in a biased population size estimate (Zwane and van der Heijden 2007). In assuming missing at random (MAR) we can impute the missing values of the partially observed covariate in the other register and use this covariate to replace the strict independence assumption with independence conditional on covariates. For partially observed covariates the log-linear model is easily extendable, so that we can also conduct sensitivity analyses in this context.

We proceed as follows. In section 2 we will discuss the log-linear model for a capture-recapture model with two registers without covariates. In Section 3 we will discuss a two-register capture-recapture model and conduct a sensitivity analysis on two registers with a conditional independence. In Section 4 the independence assumption will be conditional on partially observed covariates, where a covariate has been observed in only one register. Here the sensitivity analysis is on the dependence of the partially observed covariate on the register, thus whether the covariate influences the inclusion probability of the register. Section 5 provides some extensions made to a specific model, namely for models for three registers, the multiplier method and confidence intervals.

We use two data sources to illustrate the robustness of capture-recapture methodology, which have been provided by Statistics Netherlands. We chose not to make a simulation study because researchers in the field of capture-recapture use real data and we wanted to make the impact of a possible dependence relevant to such researchers. The first data source is the GBA (Gemeentelijke Basisadministratie) which is the official Dutch Population Register containing demographic information on the 'de jure' population. The 'de jure' population differs from the 'de facto' population, the latter also containing residents who have immigrated from other countries of the European Union and did not register as such, immigrants who (are planning to) stay less than four months and illegal immigrants. An important part of the difference between the 'de jure' and the 'de facto' population is the group of temporary workers from eastern Europe, in particular Poland. The second data source is the HKS (Herkenningsdienst systeem), which is a police register of all persons suspected of known offenses. We refer the reader to van der Heijden et al. (2012) for more details on the registers.

2. Two Registers Without Covariates

The simplest population size estimation model makes use of two registers, 1 and 2. Let variables A and B respectively denote inclusion in registers 1 and 2. Let the levels of A be indexed by i $(i = 0, 1)$ where $i = 0$ stands for "not included in register 1", and $i = 1$ stands for "included in register 1". Similarly, let the levels of B be indexed by j $(j = 0, 1)$. Expected values are denoted by m_{ij}. Observed values are denoted by n_{ij} with $n_{00} = 0$, because there are no observations for the cases that belong to the population but were not present in either of the registers.

Recall that one of the assumptions in population size estimation is that the probability of being in the first register is independent of the probability of being in the second register. Under independence, the log-linear model for the counts n_{01}, n_{10} and n_{11} is:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B \tag{1}$$

where we used the identifying restrictions $\lambda_0^A = \lambda_0^B = 0$. There are two ways to derive the estimate of the missed part of the population. First, by $\hat{m}_{00} = \exp(\hat{\lambda})$, and second, by using the property that the odds ratio under independence is 1, that is, $m_{00}m_{11}/m_{10}m_{01} = 1$ so that:

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}. \tag{2}$$

For the first way of estimating the missed portion of the population we need an estimate of λ in (1). There are several ways to estimate the parameters in (1), and it suits our purposes later on to use the generalized linear model. We assume that n_{ij} follow a Poisson distribution; a log link connects the expected values m_{ij} to the linear predictor. In terms of matrices and vectors we get

$$\log \begin{pmatrix} m_{11} \\ m_{10} \\ m_{01} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^A \\ \lambda_1^B \end{pmatrix} \tag{3}$$

where the right-hand side of (3) leads to a vector with elements $\left[\lambda + \lambda_1^A + \lambda_1^B, \lambda + \lambda_1^A, \lambda + \lambda_1^B\right]$. Thus the estimates of λ, λ_1^A and λ_1^B will get us estimates

$\hat{m}_{11}, \hat{m}_{10}$ and $\hat{m}_{01}$ of which also the missed portion of the population $\hat{m}_{00}$ is found by log $(\hat{m}_{00}) = \hat{\lambda}$, so that $\hat{m}_{00} = \exp(\hat{\lambda})$.

However, the problem with using the independence model is that independence is an unverifiable assumption, that is, we can not verify independence from the data. Thus the Poisson log-linear model for independence works under the assumption that the interaction parameter $\lambda_{ij}^{AB} = 0$. As noted before, this assumption could be violated and the population size estimate under independence may well be inaccurate. We are interested in what happens to the population size estimate when we assume independence when actually the inclusion probabilities of inclusion in registers 1 and 2 are dependent.

The approach we advocate is to include a fixed interaction parameter $\tilde{\lambda}_{ij}^{AB}$ in the model, where the tilde indicates that the interaction parameter is not estimated but fixed. By choosing interesting values for $\tilde{\lambda}_{ij}^{AB}$ we can conduct a sensitivity analysis on the population size estimate. The log-linear model then becomes:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \tilde{\lambda}_{ij}^{AB} \tag{4}$$

where we used the identifying restrictions $\tilde{\lambda}_{00}^{AB} = \tilde{\lambda}_{10}^{AB} = \tilde{\lambda}_{01}^{AB} = 0$. In matrix terms we get:

$$\log \begin{pmatrix} m_{11} \\ m_{10} \\ m_{01} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^A \\ \lambda_1^B \\ \tilde{\lambda}_{11}^{AB} \end{pmatrix} \tag{5}$$

The log-linear model for independence is a special case of this saturated model when $\lambda_{ij}^{AB} = \tilde{\lambda}_{ij}^{AB} = 0$. Dependence can be introduced to log-linear models by fixing $\tilde{\lambda}_{ij}^{AB}$ to anything but 0. In software for Poisson regression, Model (4) and (5) can be fit by entering $\tilde{\lambda}_{ij}^{AB}$ as a so-called offset. When $\tilde{\lambda}_{ij}^{AB} \neq 0$, $\hat{\lambda}$ in (5) differs from $\hat{\lambda}$ in (3).

Note that interesting values for $\tilde{\lambda}_{ij}^{AB}$ can be chosen using a direct relationship between λ_{ij}^{AB} and the odds ratio θ, which is:

$$\theta = \frac{m_{11} m_{00}}{m_{10} m_{01}} = \exp \tilde{\lambda}_{11}^{AB}. \tag{6}$$

Using the Poisson log-linear model with an offset is a general approach for carrying out a sensitivity analysis. The approach is general in the sense that it can be applied in more complicated log-linear models, for example when it is desirable to investigate violations of more than one assumption simultaneously (cf. the models discussed in Subsection 4.2). For completeness we also discuss a second method that is simpler but less general.

The second way of estimating the missed portion of the population is by using odds ratios directly, as has been done in Brown et al. (1999). We show this second way to give a full overview of the method. This also provides for simpler notation, which we will use in the rest of the article. Under independence, the odds ratio $m_{11} m_{00} / m_{10} m_{01} = 1$, and by rewriting and replacing the expected values with observed values, we get maximum likelihood estimate (2). We can impute dependence by making the odds ratio $\theta \neq 1$. Thus $\theta = m_{11}m_{00}/m_{10}m_{01}$, and

$$\hat{m}_{00(\theta)} = \theta \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \theta \frac{n_{10}n_{01}}{n_{11}} = \theta \hat{m}_{00}. \tag{7}$$

Note that $\hat{m}_{00(\theta)}$ can be found simply by multiplying the estimate under independence, $\hat{m}_{00}$, with θ. Both approaches, the log-linear Poisson regression with an offset and the odds ratio, yield the same $\hat{m}_{00}$. We will use the odds ratio to denote dependence as it provides a simpler notation than the interaction parameter $\tilde{\lambda}_{ij}^{AB}$.

The methods just described allow us to study the impact of a violation of the independence assumption as a function of θ. To get the population size estimate, let n be the total of observed cases, $n = n_{01} + n_{10} + n_{11}$, let $\hat{N}$ be the population size estimated under $\theta = 1$, thus $\hat{N} = n + \hat{m}_{00}$, and define $\hat{N}_{(\theta)}$ as the estimated population size under dependence of size θ, $\hat{N}_{(\theta)} = n + \hat{m}_{00(\theta)} = n + \theta\hat{m}_{00}$. It follows that under negative dependence (i.e., $\theta < 1$), $\hat{N}$ will be an overestimation compared to $\hat{N}_{(\theta)}$, and under a positive dependence (i.e., $\theta > 1$), $\hat{N}$ will be an underestimation compared to $\hat{N}_{(\theta)}$. The bias will be smaller the closer θ is to 1.

Assume that Register 1 has a better coverage of the population than Register 2. Then when $n_{11}/(n_{11} + n_{01})$ is high the observed coverage is high, and vice versa. Brown et al. (2006) showed that as the observed coverage increases, the number of individuals that are missed by Register 1 reduces and $n_{11}/n_{10}n_{01}$ increases so that $n_{10}n_{01}/n_{11} = \hat{m}_{00}$ decreases. Then, the implied coverage of Register 1 is high, so that $\hat{m}_{00}$ is reasonably robust to dependence. When the observed coverage decreases, the number of individuals missed by Register 1 increases and $n_{11}/n_{10}n_{01}$ decreases. Then the implied coverage of Register 1 will be low, so that $\hat{m}_{00}$ is less robust to dependence.

To illustrate, we use two registers of Statistics Netherlands, the GBA and the HKS, on people with Afghan, Iranian, or Iraqi (AII) nationality living in the Netherlands in 2007 (shown in Table 1; van der Heijden et al. 2012), and on people with a Polish nationality living in the Netherlands in 2009 (shown in Table 1; van der Heijden et al. 2011).

For the people with Afghan, Iraqi, and Iranian nationality $\hat{m}_{00} = 6,170$ under independence between the registers GBA and HKS. The population size estimated under $\theta = 1$ becomes $\hat{N} = 27,594 + 6,170 = 33,764$. Then, under dependence between the registers GBA and HKS the estimated population size becomes $\hat{N}_{(\theta)} = 27,594 + (\theta^{*}6,170)$, see (7).

To investigate the robustness of the estimate under dependence we vary θ from 0.5 to 2. In the log-linear Poisson regression approach this corresponds to using offsets varying between log(0.5) and log(2). Table 2 shows $\hat{m}_{00(\theta)}$, the population size estimate $\hat{N}(\theta)$, the estimated

Table 1. The observed values for the two nationalities, with the Afghan, Iraqi, and Iranian people residing in the Netherlands in 2007 on the left, and the Polish people residing in the Netherlands in 2009 on the right.

AII	HKS		Polish	HKS	
GBA	1	0	GBA	1	0
1	1,085	26,254	1	374	39,488
0	255	-	0	1,445	-

Table 2. Sensitivity analysis of the population size estimate for the people residing in the Netherlands in 2007 with Afghan, Iraqi, and Iranian nationality (upper panel) and for people with Polish nationality in 2009 (lower panel).

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AII	$\hat{m}_{00(\theta)}$	3,085	4,114	6,170	9,255	12,341
	$\hat{N}_{(\theta)}$	30,679	31,708	33,764	36,849	39,935
	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
	se	223	293	441	647	864
Polish	$\hat{m}_{00(\theta)}$	76,284	101,712	152,567	228,851	305,135
	$\hat{N}_{(\theta)}$	117,591	143,019	193,874	270,158	346,442
	$\hat{N}/\hat{N}_{(\theta)}$	1.65	1.36	1.00	0.72	0.56
	se	4,473	6,024	8,787	13,630	17,866

relative bias $\hat{N}/\hat{N}_{(\theta)}$ and the bootstrapped standard error (se) of the estimate for both nationalities (details about the parametric bootstrap are provided in Subsection 5.3). As can be seen from the upper panel of Table 2, for the people with Afghan, Iraqi, and Iranian nationality under a dependence of $\theta = 0.5$, the estimate $\hat{m}_{00(\theta)}$ is half the size of the population size estimate under independence, and for a dependence of $\theta = 2$ the estimate $\hat{m}_{00}$ is twice the size of the population size estimate under independence. If in the population the registers are dependent with a true size θ, the population size estimate under independence varies between a ten percent overestimation and a 15 percent underestimation. Thus when the true $\theta \neq 1$ our population size estimate under independence remains fairly accurate.

However, for the Polish people the population size estimate under dependence is not robust. As can be seen from the lower panel of Table 2, if in the population the registers are dependent with a true size θ, the population size estimate under independence deviates between a 65 percent overestimation and 44 percent underestimation. Thus when the true $\theta \neq 1$, the population size estimate under independence for the Polish people is not robust.

The most important reason why the population size estimate deviates this much is because the implied coverage of the people with Afghan, Iraqi, and Iranian nationality is smaller than for the individuals with a Polish nationality. For example, 1,085 is $1,085/(1,085 + 255) = 0.81$, thus 81 percent of implied coverage of the GBA measured by the HKS. By contrast, for the individuals with Polish nationality the implied coverage of the GBA is only 21 percent, confirming the research by Brown et al. (2006) that as the observed coverage increases, the implied coverage increases and thus the population size estimate is more robust against dependence.

The estimated standard error of $\hat{N}_{(\theta)}$ is mainly determined by the size of $\hat{m}_{00(\theta)}$, and this explains the sharp rise of the standard error from $\theta = .50$ to $\theta = 2.00$ and the difference in standard error between the individuals with Afghan, Iraqi, and Iranian nationality and the individuals with Polish nationality.

3. Two Registers With Fully Observed Covariates

Covariates were first introduced to capture-recapture by Alho (1990) to reduce the heterogeneity resulting from individual differences on that covariate. As such, if covariates are available, the generally nonfeasible independence assumption can be replaced with a less strict conditional independence assumption, where independence is conditional on covariates (Bishop et al. 1975; van der Heijden et al. 2012). This assumption is less stringent because it can take into account inclusion probabilities that are heterogeneous over the levels of the included covariate. Another advantage of using covariates is that it allows us to investigate the characteristics of the missing portion of the population.

Suppose we have observed covariate X, where the levels of X are indexed by x, $(x = 0,\ 1)$. Under independence conditional on X, there are two zero counts for cases not found in either register, namely for $x = 0$ and for $x = 1$. Let m_{ijx} denote the expected values for A, B and X. The log-linear model for independence for two registers and covariate X is

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX}, \tag{8}$$

with identifying restrictions that a parameter equals zero when i or j or $x = 0$. When assuming independence between A and B conditional on X, $\lambda_{ij}^{AB} = \lambda_{ijx}^{ABX} = 0$. We use the notation of Bishop et al. (1975) to denote hierarchical log-linear models, that is, we denote this model as $[AX][BX]$.

In Section 2 we discussed two ways to estimate population sizes in a sensitivity analysis, namely one using an offset in a Poisson log-linear model and another using odds ratios directly. Here we only discuss the first way as it is more general. We assume that n_{ijx} follow a Poisson distribution and a log link connects the expected value m_{ijx} to the linear predictor.

It is important to note that in this context, too, sensitivity analyses are useful for assessing the impact of assumptions that are not verifiable from the data under study. Here conditional independence is the unverifiable assumption, since model $[AX][BX]$ is the saturated model. By contrast, model violations for more restricted models are verifiable in the data, for example for a model such as $[A][BX]$. Hence, the impact of interaction between A and X does not have to be investigated via a sensitivity analysis. However, when there may be dependence between A and B, a sensitivity analysis is useful.

We model dependence in the data by adding fixed parameters $\tilde{\lambda}_{ij}^{AB} + \tilde{\lambda}_{ijx}^{ABX}$ to Model (8). We again work under the saturated model, as the number of parameters to be estimated is equal to the number of observed parameters:

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX} + \tilde{\lambda}_{ij}^{AB} + \tilde{\lambda}_{ijx}^{ABX}, \tag{9}$$

with the additional restrictions that parameters $\tilde{\lambda}_{ij}^{AB}$ and $\tilde{\lambda}_{ijx}^{ABX}$ equal zero when i or j or $x = 0$.

Under dependence between A and B given X, the association between the odds ratio θ_x and the log-linear parameters is:

$$\theta_x = \frac{m_{11x}m_{00x}}{m_{10x}m_{01x}} = \exp\left(\tilde{\lambda}_{11}^{AB} + \tilde{\lambda}_{11x}^{ABX}\right). \tag{10}$$

When we assume that dependence for $x = 0$ is identical to dependence for $x = 1$, then:

$$\theta = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp\left(\tilde{\lambda}_{11}^{AB}\right). \tag{11}$$

Table 3. The observed values for the Afghan, Iraqi, and Iranian people, males on the left panel and females on the right panel.

Males	HKS		Females	HKS	
GBA	1	0	GBA	1	0
1	972	14,883	1	113	11,371
0	234	-	0	21	-

We estimate (9) using log-linear Poisson regression with for cell (1,1,0) the offset $\tilde{\lambda}_{11}^{AB}$ and for cell (1,1,1) the offset $\tilde{\lambda}_{11}^{AB} + \tilde{\lambda}_{111}^{ABX}$. After estimating (9), estimates for the missed portions of the population are found by $\hat{m}_{000} = \exp(\hat{\lambda})$ and $\hat{m}_{001} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^X\right)$.

Table 3 shows the data for the Afghan, Iraqi, and Iranian people distributed over males $(x = 0)$ and females $(x = 1)$. Under conditional independence, $\hat{m}_{000} = 3,583$ and $\hat{m}_{001} = 2,113$. Taken together, both registers missed 5,696 cases. Note that conditional independence does not imply marginal independence under model $[AX][BX]$, since the marginal odds ratio $1,085^{*}5,696/26,254^{*}255 = 0.92$, and hence shows dependence (under marginal independence it would be equal to 1).

We estimate the parameters in (9) with a Poisson regression with $\tilde{\lambda}_{ijx}^{ABX} = 0$, so that the odds ratio of the males equals the odds ratio of the females (cf. (11)). The upper panel of Table 5 shows the results of the sensitivity analysis for the people with Afghan, Iraqi, and Iranian nationality in 2007 and the covariate gender. If in the population the registers are dependent with a true size θ, the population size estimate under independence varies between a nine percent overestimation to a 15 percent underestimation. As $\hat{m}_{00(\theta)}$ is relatively small, the standard error is relatively small. Thus when the true $\theta = 0.5$ but we estimate under $\theta = 1$, the population size estimate under independence is fairly robust.

For the people with a Polish nationality residing in the Netherlands in 2009 the covariate gender is also used. Under conditional independence, the estimate $\hat{m}_{00x} = 144,548$. The lower panel of Table 5 shows the sensitivity analysis of the population size estimator under conditional independence. If in the population the registers are dependent with a true size θ, the population size estimate under independence ranged between a 58 percent overestimation and a 42 percent underestimation. Thus when the true $\theta \neq 1$, the population size estimate deviates greatly from the population size estimate under $\theta = 1$, indicating that for this dataset the population size estimate under independence is not robust.

We note that this example uses a covariate with only two levels. One can easily extend this to covariates with more levels. Assume covariate W has three levels, where the levels of W are indexed by w ($w = 0,\ 1,\ 2$). Then there are three zero counts, namely for $w = 0$, $w = 1$ and $w = 2$. One can estimate the zero counts using Equation (10), where estimates

Table 4. The observed values for the Polish people, males on the left panel and females on the right panel.

Males	HKS		Females	HKS	
GBA	1	0	GBA	1	0
1	313	19,152	1	61	20,336
0	1,349	-	0	96	-

Table 5. Sensitivity analysis for the people with Afghan, Iraqi, and Iranian (AII) nationality residing in the Netherlands in 2007 (upper panel), and the people with Polish nationality residing in the Netherlands in 2009 (lower panel), conditional on gender.

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AII	$\hat{m}_{00}$	2,848	3,797	5,696	8,544	11,392
	$\hat{N}_{(\theta)}$	30,442	31,391	33,290	36,138	38,986
	$\hat{N}/\hat{N}_{(\theta)}$	1.09	1.06	1.00	0.92	0.85
	Se	292	390	576	863	1144
Polish	$\hat{m}_{00}$	57,274	76,365	114,548	171,821	229,095
	$\hat{N}_{(\theta)}$	98,581	117,672	155,855	213,128	270,402
	$\hat{N}/\hat{N}_{(\theta)}$	1.58	1.32	1.00	0.73	0.58
	Se	3,814	5,088	7450	11,465	15,135

for the missed portions of the population are found by $\hat{m}_{000} = \exp(\hat{\lambda})$ and $\hat{m}_{001} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^W\right)$ and $\hat{m}_{002} = \exp\left(\hat{\lambda} + \hat{\lambda}_2^W\right)$.

4. Two Registers With Partially Observed Covariates

In Section 3 we used covariates that are present in both registers (fully observed covariates) to replace the strict independence assumption with an independence assumption conditional on covariates. However, a register usually also has a set of variables that are only measured in one register and not in the other register (partially observed covariates). Partially observed covariates in A are usually ignored because including them leads to missing data in B for those individuals that are not in A, and vice versa. When these covariates are related to the inclusion probability, ignoring the partially observed covariates can lead to a biased population size estimate (Zwane and van der Heijden 2007; van der Heijden et a1. 2012).

4.1. Partially Observed Covariates

Partially observed covariates can be approached as a missing data problem (Zwane and van der Heijden 2007). If we assume MAR mechanism for the data, then we can use the Expectation-Maximization (EM) algorithm to estimate the missing values of the partially observed covariate of register 1 (and 2) for the individuals not present in Register 1 (and 2). MAR assumes that the probability of missingness depends only on the observed variables in the capture-recapture model (Little and Rubin 1987). When the assumption of MAR has been satisfied, the EM algorithm will give unbiased estimates.

Suppose register 1 has the covariate X_1, indexed by $k(k = 0, 1)$, where the values for X_1 are missing for $A = 0$ because X_1 is not in register 2. Assume that register 2 has the covariate X_2, indexed by $l(l = 0, 1)$, where the values for X_2 are missing for $B = 0$ because X_2 is not in register 1. The log-linear conditional independence model for two registers, with two partially observed covariates X_1 and X_2, is denoted as

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2}, \qquad (12)$$

Table 6. Expected values for two registers and two partially observed covariates.

		B = 1		B = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
A = 1	$X_1 = 1$	m_{1111}	m_{1110}	m_{1011}	m_{1010}
	$X_1 = 0$	m_{1101}	m_{1100}	m_{1001}	m_{1000}
A = 0	$X_1 = 1$	m_{0111}	m_{0110}	m_{0011}	m_{0010}
	$X_1 = 0$	m_{0101}	m_{0100}	m_{0001}	m_{0000}

with identifying restrictions $\lambda_{ij}^{AB} = \lambda_{ik}^{AX_1} = \lambda_{jl}^{BX_2} = \lambda_{ijk}^{ABX_1} = \lambda_{ijl}^{ABX_2} = \lambda_{ijkl}^{ABX_1X_2} = 0$. The conditional independence model is denoted by $[AX_2][BX_1][X_1X_2]$. Inclusion of the parameter $\lambda_{il}^{AX_2}$ instead of the parameter $\lambda_{ik}^{AX_1}$ may seem counterintuitive, but no interaction for A and X_1 can be identified as the levels of X_1 do not vary over individuals for which $A = 0$, and similarly for B and X_2 (Zwane and van der Heijden 2007).

Table 6 illustrates that two registers with two covariates lead to 16 cells. However, because our covariates are only partially observed, columns $X_2 = 1$ and $X_2 = 0$ for $B = 0$ are collapsed, just as rows $X_1 = 1$ and $X_1 = 0$ for $A = 0$ are collapsed. In other words, we do not observe counts for m_{0111} and m_{0101} but only one count for the sum $m_{0111} + m_{0101}$, and similarly for $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$. Note that we have no observed values for m_{0011}, m_{0001}, m_{0010} and m_{0000}, as these refer to individuals who are in neither of the registers. Thus model $[AX_2][BX_1][X_1X_2]$ is saturated with eight observed values and eight parameters to be estimated.

Using the EM algorithm we first estimate the four missing cells, that is, the cells that are missing because the covariates are only partially observed. In the E-step we spread out the four sums $m_{0111} + m_{0101}$, $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$ over the eight cells to get an expectation for the missing data. In the M-step we estimate log-linear model (12) to the completed table of twelve cells. For estimation, we assume that the twelve counts follow a Poisson distribution and a log link connects the expected counts to the linear predictor. The resulting estimates are then used for the E-step where in the M-step, following (12), we estimate the parameters again.

To illustrate we once more use the data on the people with Afghan, Iraqi, and Iranian nationality residing in the Netherlands in 2007 with two partially observed covariates (van der Heijden et al. 2012). The GBA has the partially observed covariate marital status (X_1), where $X_1 = 1$ denotes either being married or living together and $X_1 = 0$ denotes either unmarried, divorced or widowed. The HKS has the partially observed covariate police region (X_2), where $X_2 = 1$ denotes residing in one of the five biggest cities of the Netherlands (i.e., Amsterdam, Rotterdam, Utrecht, The Hague, and Eindhoven) and $X_2 = 0$ denotes residing in the rest of the country.

Due to the log-linear model used, the first four observed values remain unchanged for each iteration (for $GBA = 1$ and $HKS = 1$). The upper panel of Table 7 shows the observed counts and the lower panel of Table 7 shows the fitted counts after convergence of the EM algorithm. As an example, the observed value of 91 (for $X_2 = 1$, where X_1 values are missing under $GBA = 0$) is spread out into the values 64 for $X_1 = 1$ and 27 for $X_1 = 0$. After convergence, the unobserved part of the population is estimated. In total,

Table 7. Data for the Afghan, Iraqi, and Iranian people residing in the Netherlands in 2007, spread out over the partially observed covariates marital status X_1 and police region X_2

Panel 1: The observed counts

		HKS = 1		HKS = 0
		$X_2 = 1$	$X_2 = 0$	X_2 missing
GBA = 1	$X_1 = 1$	259	539	13,898
	$X_1 = 0$	110	177	12,356
GBA = 0	X_1 missing	91	164	-

Panel 2: The fitted frequencies

		HKS = 1		HKS = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
GBA = 1	$X_1 = 1$	259	539	4,511	9,387
	$X_1 = 0$	110	177	4,736	7,620
GBA = 0	$X_1 = 1$	64	123	1,112	2,150
	$X_1 = 0$	27	41	1,168	1,745

we estimate that there were 33,770 individuals with Afghan, Iraqi, and Iranian nationality residing in the Netherlands in 2007.

4.2. *Sensitivity Analyses*

We again make use of a sensitivity analysis to investigate the unverifiable assumption of independence conditional on partially observed covariates. Model violations for more restricted models are verifiable in the data. For example, using a model such as $[AX_2][BX_1]$ allows us to investigate absence of interaction $\lambda_{kl}^{X_1X_2}$ in the data. Thus the impact of an interaction between X_1 and X_2 does not need to be investigated via a sensitivity analysis. However, in this context (12) is the saturated model and therefore model violations such as dependence between A and X_1, between B and X_2, and between A and B are unverifiable, rendering it useful to conduct a sensitivity analysis. Note that in the previous sections we used a sensitivity analysis to assess the interaction between the two registers. In this section we assess not only the interaction between A and B, but also the interaction between the register and its partially observed covariate. To exemplify, we introduce an interaction parameter that simulates dependence between the GBA and marital status. Such a dependence would imply that marital status influences the inclusion probability of being in the GBA.

The log-linear model for an interaction between A and B would be:

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{ij}^{AB}, \quad (13)$$

with additional identifying restrictions that $\tilde{\lambda}_{ij}^{AB} = 0$ when i or j equals 0. Here $\exp\left(\tilde{\lambda}_{ij}^{AB}\right)$ is the conditional odds ratio for the interaction between A and B.

Assume the partially observed covariate marital status is related to the inclusion probability of the GBA, thus $\lambda_{ik}^{AX_1} \neq 0$. Because the interaction between A and X_1 is unverifiable from the data, the fixed parameter $\tilde{\lambda}_{ik}^{AX_1}$ has been added to the log-linear model (12). We continue to work under the saturated model:

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{ik}^{AX_1}, \qquad (14)$$

with additional identifying restrictions that $\tilde{\lambda}_{ik}^{AX_1} = 0$ when i or k equals 0. The same can be done for the interaction between B and X_2. When the partially observed covariate X_2 is related to the inclusion probability of register B, $\lambda_{jl}^{BX_2} \neq 0$. We add fixed parameter $\tilde{\lambda}_{jl}^{BX_2}$ to the log-linear model. The log-linear model then becomes:

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{jl}^{BX_2}, \qquad (15)$$

with additional identifying restrictions that $\tilde{\lambda}_{jl}^{BX_2} = 0$ when j or l equals 0. We can estimate (13), (14) and (15) via Poisson regressions with offsets. Note that in modeling these relationships we have to fix the offset variable on a log scale. Then we can estimate the portions of the population that both registers have missed by $\hat{m}_{0000} = \exp(\hat{\lambda})$, $\hat{m}_{0010} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^{X_1}\right)$, $\hat{m}_{0001} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^{X_2}\right)$ and $\hat{m}_{0011} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^{X_1} + \hat{\lambda}_1^{X_2} + \hat{\lambda}_{11}^{X_1X_2}\right)$.

The upper panel of Table 8 shows the sensitivity analysis for the interaction between A and B, the middle panel shows the sensitivity analysis for the interaction between A and X_1 and the lower panel shows the sensitivity analysis for the interaction between B and X_2 for the Afghan, Iraqi, and Iranian people. As can be seen, for the interaction between A and B, the relative bias is similar to the bias found in Tables 2 and 5. If in the population the GBA and marital status are dependent with a true size θ, the estimation under independence deviates between a 2.22 percent overestimation to a 2.89 percent underestimation, and the estimation under independence between the HKS and police region deviates between a 0.23 percent underestimation and a 0.19 percent overestimation. Thus for the interactions AX_1 and BX_2, when the true $\theta \neq 1$, the population size estimate under independence remains fairly robust.

We have done the same for the people with Polish nationality residing in the Netherlands in 2009. The observed values are shown in the upper panel of Table 9 and the expected

Table 8. Sensitivity analysis of the population size estimate for the people residing in the Netherlands in 2007 with an Afghan, Iraqi, and Iranian nationality with the interaction A and X_1 (upper panel) and the interaction between B and X_2 (lower panel).

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AB	$\hat{m}_{00(\theta)}$	3.088	4,117	6,176	9,264	12,352
	$\hat{N}_{(\theta)}$	30.682	31,711	33,770	36,858	39,946
	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
AX1	$\hat{m}_{00(\theta)}$	5,443	5,711	6,176	6,736	7,179
	$\hat{N}_{(\theta)}$	33,037	33,305	33,770	34,330	34,773
	$\hat{N}/\hat{N}_{(\theta)}$	1.0222	1.0140	1.00	0.9837	0.9711
BX2	$\hat{m}_{00(\theta)}$	6,253	6,220	6,176	6,136	6,112
	$\hat{N}_{(\theta)}$	33,847	33,814	33,770	33,730	33,706
	$\hat{N}/\hat{N}_{(\theta)}$	0.9977	0.9987	1.00	1.0012	1.0019

Table 9. The observed counts for the people with Polish nationality residing in the Netherlands in 2009 (upper panel) and the fitted frequencies spread out over the partially observed covariates (lower panel).

Panel 1: The observed counts

		HKS = 1		HKS = 0
		$X_2 = 1$	$X_2 = 0$	X_2 missing
GBA = 1	$X_1 = 1$	111	188	25,416
	$X_1 = 2$	32	43	14,072
GBA = 0	$X_1 = 1$	603	842	

Panel 2: The fitted frequencies

		HKS = 1		HKS = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
GBA = 1	$X_1 = 1$	111	188	9,435	15,981
	$X_1 = 2$	32	43	6,004	8,068
GBA = 0	$X_1 = 1$	468	685	39,787	58,250
	$X_1 = 2$	135	157	25,318	29,408

frequencies are shown in the lower panel of Table 9. Again a sensitivity analysis has been conducted, which is shown in Table 10. Just as with the individuals with Afghan, Iraqi, and Iranian nationality, the estimates and thus the relative bias under dependence between *A* and *B* remains unchanged (cf. Tables 2 and 5). If in the population the GBA and marital status are dependent with a true size θ, the population size estimate under independence ranges from a seven percent overestimation to a nine percent underestimation (upper panel). The estimate under independence between the HKS and police region deviates from a two percent underestimation to a two percent overestimation (lower panel). Thus when the true $\theta \neq 1$, the population size estimate under independence remains fairly robust.

Table 10. Sensitivity analysis of the population size estimate for the the people residing in the Netherlands in 2009 with Polish nationality with the interaction between A and X_1 (upper panel) and the interaction between B and X_2 (lower panel).

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AB	$\hat{m}_{00(\theta)}$	76,381	101,842	152,762	229,143	305,524
	$\hat{N}_{(\theta)}$	117,688	143,149	194,069	270,450	346,832
	$\hat{N}/\hat{N}_{(\theta)}$	1.65	1.36	1.00	0.71	0.56
AX1	$\hat{m}_{00(\theta)}$	139,494	144,238	152,762	163,584	172,582
	$\hat{N}_{(\theta)}$	180,801	185,545	194,069	204,891	213,889
	$\hat{N}/\hat{N}_{(\theta)}$	1.07	1.05	1.00	0.95	0.91
BX2	$\hat{m}_{00(\theta)}$	156,616	155,004	152,762	150,707	149,429
	$\hat{N}_{(\theta)}$	197,923	196,311	194,069	192,014	190,736
	$\hat{N}/\hat{N}_{(\theta)}$	0.98	0.99	1.00	1.01	1.02

Under the use of partially observed covariates it becomes clear why the log-linear Poisson regression provides a more general approach than using odds ratios to implement the sensitivity analyses. When using log-linear Poisson regression the process becomes vastly simpler, in that the offset can be set to any number per cell. When multiple different offsets are in use, the log-linear Poisson regression allows for this complexity, whereas implementing odds ratios may become gruesome.

5. Miscellany

5.1. Extension to Multiple Sources

One way to make the impact of possible violations of the independence assumption less severe is by conditioning on covariates, as we have seen in Section 3 and 4. Another way to make the impact of possible violations of the independence assumption less severe is by adding registers, when more registers are available (cf. Baffour et al. 2013). Assume we have three registers 1, 2 and 3, where the variables A, B and C respectively stand for inclusion in the registers. We denote the expected values m_{ijp} where i, j, $p = 1$ stand for the inclusion into Registers 1, 2 and 3 respectively and where i, j, $p = 0$ stands for the absence in registers 1, 2 and 3.

For three variables, the saturated log-linear model is denoted by

$$\log\ m_{ijp} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_p^C + \lambda_{ij}^{AB} + \lambda_{ip}^{AC} + \lambda_{jp}^{BC}, \tag{16}$$

with identifying restrictions that a parameter equals zero when i, j or $p = 0$. We assume that interaction parameter $\lambda_{ijp}^{ABC} = 0$. Model $[AB][BC][AC]$ is the saturated model, as the number of observed parameters equals the number of parameters to be estimated. With d registers, we assume that the d-factor interaction is absent.

For estimation, assume that n_{ijp} follow a Poisson distribution and a log link connects the expected value m_{ijp} to the linear predictor. We can estimate the parameters in (16) via a Poisson log-linear regression.

Model $[AB][BC][AC]$ assumes that odds conditional on a third variable are equal, for example for the odds ratio between A and B given C we find

$$\frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}}. \tag{17}$$

Model (16) assumes that for estimation with odds ratios under saturated model $[AB][BC][AC]$ we get:

$$\frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} = \hat{m}_{000}. \tag{18}$$

An estimate for $\hat{m}_{000}$ is easily derived from (17) as $[AB][AC][BC]$ is the saturated model in this context; absence of the three-factor interaction is an unverifiable assumption as it cannot be verified in the data. More restricted models such as $[AB][AC]$ are verifiable in the data. However, we can investigate the robustness of the population size estimate against violations of the assumption that the three-factor interaction is absent by fixing the interaction parameter to anything but 0, that is, $\tilde{\lambda}_{ijp}^{ABC} \neq 0$. Thus the log-linear model becomes:

$$\log\ m_{ijp} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_p^C + \lambda_{ij}^{AB} + \lambda_{ip}^{AC} + \lambda_{jp}^{BC} + \tilde{\lambda}_{ijp}^{ABC}, \tag{19}$$

with the additional identifying restriction where parameter $\tilde{\lambda}_{ijp}^{ABC}$ equals zero when i or j or $p = 0$. The population size estimate under (19) can be estimated using Poisson log-linear regression with parameter $\tilde{\lambda}_{ijp}^{ABC}$ as an offset.

Under dependence between A and B given C, the association between the odds ratio θ and the log-linear parameters is:

$$\theta_{AB}^{(p=0)} = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \exp\ (\lambda_{11}^{AB}), \tag{20}$$

and:

$$\theta_{AB}^{(p=1)} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp\ (\lambda_{11}^{AB} + \lambda_{111}^{ABC}). \tag{21}$$

When we assume that the odds ratio between A and B is the same for $p = 0$ and $p = 1$, we get

$$\theta_{AB} = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp\ (\lambda_{11}^{AB}). \tag{22}$$

When more registers are available we can use these extra registers to reduce the impact of violations of the independence assumption. As we have shown, the log-linear model is easily generalizable to multiple registers.

5.2. *Multiplier Method*

The multiplier method is an alternative method to estimate the size of a population and it is used, amongst others, in drug use research and HIV prevalence (European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) 1997; Cruts and van Laar 2010; Temurhan et al. 2011). Multiplier methods are user-friendly for their mathematical simplicity, and absence of linkage, and are straightforward to use. At least two data sources are needed to use the multiplier method, usually a comprehensive register and a survey. For example, assume we wish to estimate the number of Polish people residing in the Netherlands in 2013. We assume that everyone has an equal chance of going to a hospital, thus we go to hospitals to assess how many Polish patients there are, and ask them whether they are in the GBA. Then assume the data we found is the data from Table 11. There are 200 Polish people, of which 150 are in the GBA. Thus p(GBA |Hospital) $= 0.75$. If a total of 40,000 Polish people are registered, in the GBA, this means our actual total should be $40,000/0.75 = 53,333$ and we missed $53,333 - 40,000 = 13,333$ people who are not registered in the GBA.

The multiplier method can also be explained from the perspective of capture-recapture methods. Using the counts provided above, we have n_{11}, n_{01} and n_{1+} so that $n_{1+} - n_{11} = n_{10}$ and Equation (2) gives $\left(39,850^{*}50\right)/150 = 13,283$. Then $\hat{N} = 150 + 50 + 39,850 +\ 13,283 = 53,333$, which is the exact same value as we got above. A sensitivity analysis could be conducted using Equation (7).

Table 11. Artificial observed data for the Polish people in the hospital

		Hospital		
		1	0	
GBA	1	150	39,850	40,000
	0	50	-	-
		200	-	-

The attractiveness of the multiplier method lies in the absence of the linkage of two sources. When estimating hidden or hard-to-reach populations, it is likely that it is difficult to obtain identifying variables to link the individuals in the samples. The absence of linkage is what makes the multiplier method different from capture-recapture. However, it has to be kept in mind that the multiplier method also relies on the underlying assumptions that being in the hospital is statistically independent from being in the GBA, and that it relies on individuals reporting their GBA status accurately when being admitted to the hospital.

5.3. *Confidence Intervals*

Apart from robustness, another aspect of the usefulness of a point estimate is its confidence interval. Parametric bootstrap confidence intervals can be used to find these confidence intervals in a simple way when dealing with incomplete contingency tables. In a parametric bootstrap sample, the estimate $\hat{m}_{00(\theta)}$ for cell (0, 0) is used in the multinomial probabilities. So for Table 1, the four probabilities are $n_{11}/\hat{N}_{(\theta)}$, $n_{10}/\hat{N}_{(\theta)}$, $n_{01}/\hat{N}_{(\theta)}$ and $\hat{m}_{00(\theta)}/\hat{N}_{(\theta)}$. A sample with size $\tilde{\lambda}_{ij}^{AB}$ is drawn with replacement. This yields four counts $n_{11}^{b=1}$, $n_{01}^{b=1}$, $n_{10}^{b=1}$ and $n_{00}^{b=1}$. The first bootstrap population size estimate $\hat{N}^{b=1}$ is found using only $n_{11}^{b=1}$, $n_{01}^{b=1}$, $n_{10}^{b=1}$, that is, ignoring $n_{00}^{b=1}$, and estimating $\hat{m}_{00(\theta)}^{b=1}$. This is repeated 10,000 times, yielding 10,000 bootstrap population size estimates. From these, 2.5 and 97.5 percentile scores are derived.

To exemplify we constructed a parametric bootstrapping confidence interval on the data presented in Section 2, which can be found in Table 12. The R code for the parametric bootstrap confidence interval can be found in Appendix A.3.

To compare, we also constructed the asymptotic confidence estimate $CI = \hat{m}_{00} + / - z_{(.975)}\left(\sqrt{\hat{\mathrm{Var}}(n)}\right)$, where $\hat{\mathrm{Var}}(n) = \left(n_{1+}n_{+1}n_{10}n_{01}\right)/\left((n_{11})^{3}\right)$ (Bishop et al. 1975). The estimated confidence interval for the Afghan, Iraqi, and Iranian people under independence is $32,905.44 - 34,623.16$, which is close to the bootstrapped confidence interval.

Table 12. Confidence intervals

Odds Ratio	AII	Polish
0.50	30,254 – 31,132	109,529 – 127,022
0.67	31,156 – 32,288	132,278 – 155,837
1.00	32,931 – 34,654	177,476 – 212,431
1.50	35,607 – 38,125	245,439 – 298,960
2.00	38,292 – 41,682	314,212 – 384,579

6. Discussion

We have shown for two different datasets that the population size estimate under dependence could be fairly robust as well as not robust at all. Deviations from independence when implied coverage is low (and thus $\hat{m}_{00}$ is high) result in bigger deviations from the population size estimate under fixed dependence than when the implied coverage is higher. Thus the estimate becomes less robust and this makes the situation worse. For the Afghan, Iraqi, and Iranian people the population size estimate did not change much when dependence was introduced; it also remained fairly robust whether or not we assumed conditional independence on fully observed covariates. However, for the Polish people, the implied coverage is small, resulting in a higher $\hat{m}_{00}$ so that the deviation from independence will be large. The resulting lack of robustness makes it even worse. Not only did the population size estimate under independence change dramatically under fixed dependence, adding a covariate to replace the strict independence assumption with the less strict independence assumption conditional on covariates changed the population size estimate but did not improve the robustness.

This reflects the fact that Polish people, much more than people from Afghanistan, Iraq, and Iran, are in the position that they work on a temporary basis without living permanently in the Netherlands. By law, it is permitted for people from European Union countries like Poland to work in the Netherlands without a work and living permit. This is not the case for people from Afghanistan, Iraq, and Iran. Therefore, the coverage of the GBA differs between both nationalities, which gives a relatively high estimation of the missed population of the Polish people compared to the Afghan, Iraqi, and Iranian people. Additionally, because we multiply $\hat{m}_{00}$ with θ, it follows that a bigger $\hat{m}_{00}$ will result in a bigger $\hat{m}_{00\theta}$ than a smaller $\hat{m}_{00}$ would when multiplied with the same θ.

We also showed how to investigate robustness of the population size estimate in models with partially observed covariates. For the example we used, the population size estimate was relatively insensitive to violation of specific conditional independence assumptions. Since adding covariates reduces heterogeneity and gives the opportunity to assess how the population is divided over the levels of the covariate, it is useful to include a partially observed covariate.

In this article we assumed that the only assumption that was violated was the independence assumption. However, violation of other assumptions could also have a large impact on the population size estimate. In particular, research on violation of the assumptions that the registers are perfectly linked as well as that the population is closed during the observation period is needed to draw conclusions on the usefulness of the capture-recapture method for estimating the undercoverage of census data.

We have chosen a range of odds ratio from 0.5 to 2. To our knowledge, it is not possible to get an accurate estimation of what a realistic θ value would be, since it is impossible to ascertain θ from the data. One way of dealing with the strict independence assumption is by adding a third register, hence using another source to estimate θ, as has been done by Brown et al. (2006) who created an adjustment factor based on a third source for the census.

In conclusion, it is important to assess the size of the implied coverage of one of the registers. We have shown that lack of robustness under dependence is easily established when implied coverage is low. However, when implied coverage is high the population size estimate remains fairly robust. Thus, instead of accepting the population size estimate as it is, researchers should report on the robustness of their estimate.

7. Appendix

To estimate the population size under log-linear models, we have used Poisson regression with an offset in SPSS and R.

A.1. R Code

Below is given the R code to get estimates $\hat{m}_{00kl}$ in the EM algorithm, for the Polish data only.

```
##Give the data
data = c(111,188,32,43,12708,12708,7036,7036,301.5,421,301.5,421) ## Polish data
data = data*10000
freqitx = freqit1 = data

## Design matrix
A = c(1,1,1,1,1,1,1,1,0,0,0,0)
B = c(1,1,1,1,0,0,0,0,1,1,1,1)
X1 = c(1,1,0,0,1,1,0,0,1,1,0,0)
X2 = c(1,0,1,0,1,0,1,0,1,0,1,0)

## OR for independence
offst = c(0,0,0,0,0,0,0,0,0,0,0,0)
for (i in 1:50000){
glm = glm(freqitx ~ A*X2 + B*X1 + X1*X2, offset=offst, family=poisson)
freqdata = c(data[1:4])
freqfit = glm$fitted.values[5:12]
freqitx = c(freqdata,freqfit)
freqitx = round(freqitx)}

## Parameter estimates under independence
par = glm$coefficients
m0011 = as.numeric(exp(par[1]+par[3]+par[5]+par[8]))
m0010 = as.numeric(exp(par[1]+par[5]))
m0001 = as.numeric(exp(par[1]+par[3]))
m0000 = as.numeric(exp(par[1]))
matrix = matrix(c(glm$fitted.values[1],glm$fitted.values[2],
glm$fitted.values[5],glm$fitted.values[6],glm$fitted.values[3],glm$fitted.values[4],
glm$fitted.values[7], glm$fitted.values[8], glm$fitted.values[9], glm$fitted.values[10],
m0011,m0010,glm$fitted.values[11],glm$fitted.values[12],m0001,m0000),4,4,byrow
= TRUE)
N = sum(matrix)
```

```
## Define the offsets. Here we only give an example for the offsets of BX2 = 0.5
offst1 = c(-0.6931472,0,-0.6931472,0,0,0,0,0,-0.6931472,0,-0.6931472,0)
## Iterative GLM Loop for the EM algorithm
for (i in 1:50000){
glm = glm(freqitx ~ A*X2 + B*X1 + X1*X2, offset= offst1, family=poisson)
freqdata = c(data[1:4])
freqfit = glm$fitted.values[5:12]
freqitx = c(freqdata,freqfit)
freqitx = round(freqitx)}

## Calculation of estimated missed frequencies
par = glm$coefficients
m0011 = as.numeric(exp(par[1] + par[3] + par[5] + par[8]))
m0010 = as.numeric(exp(par[1] + par[5]))
m0001 = as.numeric(exp(par[1] + par[3]))
m0000 = as.numeric(exp(par[1]))

m00comp = m0011 + m0010 + m0001 + m0000
PSE = sum(data)+ m00comp
print(m00comp)
print(sum(data)+ m00comp)
print(N/PSE)
```

A.2. SPSS Syntax

```
compute freqitx = freqit1.
compute freqitx = rnd(freqitx).
execute.
DEFINE EM_PGLM()
!DO !I = 1 !TO 10000.
GENLIN freqitx BY A B X1 X2 (ORDER = ASCENDING)
/MODEL A B X1 X2 A*X2 B*X1 X1*X2 INTERCEPT = YES OFFSET = offst05
DISTRIBUTION = POISSON LINK = LOG
/SAVE MEANPRED (pred_val).
compute diff = ABS(freqit1-pred_val).
means diff.
compute freqitx = pred_val.

IF((A = 1)&(B = 1)&(X1 = 1)&(X2 = 1))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 2)&(X2 = 1))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 1)&(X2 = 2))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 2)&(X2 = 2))freqitx = freqit1.

COMPUTE freqitx = rnd(freqitx).
execute.
delete variables pred_val.
```

```
!DOEND
!ENDDEFINE.
##run the macro
EM_PGLM.
```

A.3. R Code Parametic Bootstrap

The R code presented below represents the parametric bootstrap for the Polish data from Table 1

```
data = c(374, 39488, 1445) ## Polish data
theta = 2
m00 = (data[2]*data[3])/data[1]
m00theta = m00*theta
datacomp = sum(data,m00theta)
## The estimate of N, under an offset theta
n = sum(data)
N = n + m00theta
##The relative bias under an offset theta
(n + m00)/N
## Parametric bootstrap
NN = c(N)
p = matrix(c(data/datacomp, m00theta/datacomp),1)
set.seed(N)
library(combinat)
databoot = rmultinomial(rep(NN, 10000),p)
m00boot = theta* (databoot[,2]*databoot[,3])/databoot[,1]
nboot = databoot[,1:3]
Nboot = m00boot + nboot[,1] + nboot[,2] + nboot[,3]
quantile(Nboot, c(0.025, 0.5, 0.975), type = 1)
sd = function(x) sqrt(var(x))
sd(Nboot)
```

8. References

Alho, J.M. 1990. “Logistic Regression in Capture-recapture Models.” *Biometrics* 46: 623–635. Doi: http://dx.doi.org/10.2307/2532083.

Baffour, B., J.J. Brown, and P.W.F. Smith. 2013. “An Investigation of Triple System Estimators in Censuses.” *Statistical Journal of the International Association for Official Statistics* 29: 53–68. Doi: http://dx.doi.org/10.3233/SJI-130760.

Bell, W.R. 1993. “Using Information from Demographic Analysis in Post-enumeration Survey Estimation.” *Journal of the American Statistical Association* 88: 1106–1118. Doi: http://dx.doi.org/10.1080/01621459.1993.10476381.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Brown, J.J., O. Abbott, and I.D. Diamond. 2006. "Dependence in the 2001 One-number Census Project." *Journal of the Royal Statistical Society Series A* 169: 883–902.

Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. "A Methodological Strategy for a One-number Census in the UK." *Journal of the Royal Statistical Society Series A* 162: 247–267.

Chao, A., P.K. Tsay, S.-H. Lin, W.-Y. Shau, and D.-Y. Chao. 2001. "Tutorial in Biostatistics. The Application of Capture-recapture Models of Epidemiological Data." *Statistics in Medicine* 20: 3123–3157. Doi: http://dx.doi.org/10.1002/sim.996.

Cormack, R.M. 1989. "Log-linear Models for Capture-recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Cormack, R.M., Y.-F. Chang, and G.S. Smith. 2000. "Estimating Deaths from Industrial Injury by Capture-recapture: A Cautionary Tale." *International Journal of Epidemiology* 29: 1053–1059. Doi: http://dx.doi.org/10.1093/ije/29.6.1053.

Cruts, A.A.N., and M.W. van Laar. 2010. *Aantal Problematische Harddrugsgebruikers in Nederland*. Utrecht: Trimbos Instituut.

European Monitoring Centre for Drugs and Drug Addiction (EMCDDA). 1997. *Methodological Pilot Study of Local Level Prevalence Estimates*. Lisbon: EMCDDA.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 409–439. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Hook, E.B., and R.R. Regal. 1992. "The Value of Capture-recapture Methods Even for Apparent Exhaustive Surveys." *American Journal of Epidemiology* 135: 1060–1067.

Hook, E.B., and R.R. Regal. 1995. "Capture-recapture Methods in Epidemiology: Methods and Limitations." *Epidemiologic Reviews* 17: 243–264.

Hook, E.B., and R.R. Regal. 1997. "Validity of Methods for Model Selection. *Weighting for Model Uncertainty, and Small Sample Adjustment in Capture-recapture Estimation." American Journal of Epidemiology* 145: 1138–1144. Available at: http://aje.oxfordjournals.org/content/145/12/1138.full.pdf (accessed July 2015).

Hook, E.B., and R.R. Regal. 2000. "Accuracy of Alternative Approaches to Capture-recapture Estimates of Disease Frequency: Internal Validity Analysis of Data from Five Sources." *American Journal of Epidemiology* 152: 771–779. Doi: http://dx.doi.org/10.1093/aje/152.8.771.

International Working Group for Disease Monitoring and Forecasting. 1995. "Capture-recapture and Multiple-record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7485050 (accessed July 2015).

Little, R.J., and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley and Sons.

Nirel, R., and H. Glickman. 2009. "Sample Surveys and Censuses." *Sample surveys: Design Methods and Applications* 29A: 539–565.

Office of National Statistics (ONS). 2012. *The 2011 Census Coverage Assessment and Adjustment Process*. London: Office for National Statistics.

Seber, G.A.F. 1982. *The Estimation of Animal Abundance and Related Parameters*. London: Edward Arnold.

Temurhan, M., R. Meijer, S. Choenni, M. van Ooyen-Houben, G. Cruts, and M. van Laar. 2011. "Capture-recapture Method for Estimating the Number of Problem Drug Users: The Case of the Netherlands." *In Proceedings of the Intelligence and Security Informatics Conference (EISIC)*, 12–14 september, 2011, 46–51. Available at: http://www.computer.org/csdl/proceedings/eisic/2011/4406/00/4406a046-abs.html (accessed July 2015).

Van der Heijden, P.G.M., M.J.L.F. Cruy, and G. van Gils. 2011. Aantallen Geregistreerde en Nietgeregistreerde Burgers uit MOE-landen die in Nederland Verblijven. Rapportage Schattingen 2008 en 2009. The Number of Registered and Non-registered Citizens from MOE-countries Residing in the Netherlands. Reporting Estimations 2008 and 2009. *The Hague: Ministry of Social Affairs and Employment*. Available at: http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2013/01/14/aantallen-geregistreerde-en-niet-geregistreerde-burgers-uit-moe-landen-die-in-nederland-verblijven.html [in Dutch] (accessed July 2015).

Van der Heijden, P.G.M., J. Whittaker, M.J.L.F. Cruy, 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. Doi: http://dx.doi.org/10.1214/12-AOAS536.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zwane, E.N., and P.G.M. van der Heijden. 2004. "Semiparametric Models for Capture-recapture Studies with Covariates." *Computational Statistics and Data Analysis* 47: 729–743. Doi: http://dx.doi.org/10.1016/j.csda.2003.11.010.

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. Doi: http://dx.doi.org/10.1002/sim.2577.

Received December 2013
Revised October 2014
Accepted October 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 381–396, http://dx.doi.org/10.1515/JOS-2015-0023

On Modelling Register Coverage Errors

Li-Chun Zhang[1]

Register data that originate from administrative or other secondary sources are increasingly being used to generate statistical outputs directly. The coverage of the input datasets is an important issue in this respect. Traditionally capture-recapture models have been used to deal with multiple list enumerations subjected to undercoverage errors. The aim of this article is to scope possible approaches to modelling capture-recapture data with *additional* overcoverage error. Attention is primarily given to model interpretations and conditions under which a model may provide a plausible basis for estimation and uncertainty evaluation. The setting with two list enumerations is examined in depth as it is the most common in practice. Models that can be extended to include more than two lists are identified. An additional independent coverage survey with *only* undercoverage error is always needed for estimation. Potential application to census coverage-error adjustment is discussed.

Key words: List error and catch; log-linear model; pseudoconditional independence.

1. Introduction

More and more often, register data that originate from administrative or other secondary sources are being used to generate statistical outputs directly, instead of merely supplying auxiliary information for sample surveys and census. The recent round of census provides examples of this development in a number of European countries. The coverage of the input registers has a direct bearing on the population size statistics and, in the next instance, statistics about the various characteristics of interest (Zhang 2012).

A register has undercoverage of the target population if there exist population units that are not listed in the register; it has overcoverage if not all the units in the register belong to the target population. Capture-recapture (CR) models for population size estimation (e.g., Fienberg 1972; Cormack 1989; IWGDMF 1995a and 1995b) can be used to deal with the *undercoverage* errors that exist in multiple registers. A notable application is census underenumeration adjustment using an independent U-sample coverage survey to generate recapture data. See for example Wolter (1986), Hogan (1993), Brown et al. (2011), Renaud (2007), and Nirel and Glickman (2009). Note that the term *list* (e.g., Wolter 1986) is more natural than *register* in this context, as well as in a number of situations outside official statistics, such as sizing of wildlife, hard-to-reach or clandestine populations. The two terms list and register will be used interchangeably in this article.

[1] Department of Social Statistics and Demography/S3RI, University of Southampton, Highfield, Southampton SO17 1BJ, UK Email: L.Zhang@soton.ac.uk. and Statistics Norway, Akersveien 26, PB 8131 Dep, Oslo 2213, Norway Email: lcz@ssb.no.

Acknowledgment: I thank James Brown, Peter van der Heijden, two referees and the Associate Editors for stimulating discussion, healthy criticism as well as encouragement.

When it comes to overcoverage, the standard census adjustment approach is to deploy a separate *O-sample*, selected from the census reports, to directly estimate the overcoverage rate. No explicit statistical model is applied to the O-sample, in contrast to the U-sample. Moreover, fieldwork for the O-sample can be limited or totally absent – see for example Renaud (2007) for an account of the Swiss census. On the one hand, this helps to bring down the cost; on the other hand, spurious coverage errors such as duplicate reports and misreports of census residence area can to a large extent be assessed based on record matching and clerical checks without any fieldwork. However, the ability to detect *erroneous* enumeration, that is, reports of nonexistent or out-of-scope cases, may be reduced as a result.

A modelling approach to include both under- and overcoverage errors can thus have direct relevance to the census methodology. It may potentially provide a means to assess as well as to adjust for erroneous census enumerations, provided additional register enumerations from secondary sources. For example, the Office for National Statistics in the UK is currently investigating the use of administrative data for the future provision of population statistics (ONS 2013). The same goes for those countries where the traditional census enumeration has already been replaced by population registers (e.g., Israel, Switzerland), but the O-sample deploys only limited fieldwork or no fieldwork at all.

Moreover, applications to CR data in a range of situations can be conceived. For instance, the target population may be clandestine and dynamic, such as active drug users. Relevant lists may be available from the police, clinics, and various nongovernmental organisations. Erroneous enumeration can occur in all these lists. Or, consider multiple screening procedures, each generating a list of the units with a positive test result. Only the test-positive units are subjected to a comprehensive examination, which may reveal both erroneous enumerations and underenumerations in each list. A model for predicting the errors of each test as well as the combined test results may then be of interest.

In the sequels we investigate some possible approaches to modelling two-list CR data in the presence of both over- and undercoverage errors. Section 2 briefly sets out the CR model underlying the dual-system estimator (DSE) in use for census undercoverage, as expounded in Wolter (1986). The modelling approach is extended to include the overcoverage error in Section 3. All possible standard log-linear modelling alternatives for crossclassified counts are examined, as well as an approach based on the concept of pseudoconditional independence. The emphasis is on the modelling strategy, the interpretation and the conditions under which a model may provide a plausible basis for statistical estimation and uncertainty evaluation. Models that can readily be generalised to include more than two lists are identified. In Section 4 the different models are compared to each other, using artificial CR datasets that seem relevant for the setting of census population size estimation with additional administrative register data. Discussions will be given in Section 5 regarding the future work that is needed to establish a viable estimation methodology for the census or census-like population statistics.

2. Homogeneity Model for Dual-System Estimation

Wolter (1986) discussed several CR models for census undercoverage errors. The *homogeneity* model described below underpins the DSE currently in use in a number of countries. References to the assumptions as stated by Wolter are cited and given in parentheses.

Let target population U be of unknown size N. Let A and B be two lists, both of which aim to enumerate U. Let the probability that a unit in U belongs to a particular *list domain* be given as below:

		List B		
		in	out	
List A	in	p_{11}	p_{10}	p_{1+}
	out	p_{01}	p_{00}	p_{0+}
		p_{+1}	p_{+0}	1

Each unit is assumed to follow independently (“Autonomous Independence”) the multinomial distribution (“Multinomial”) with probability p_{ab} for being included in the list domain (a, b), for $a, b = 1, 0, +$. Note that U_{00} refers to the units that are neither enumerated in A nor B. Let the list-domain size N_{ab} be observed except for N_{00} and $N = N_{++}$, that is, the matching of list A and B is error free (“Matching”). All the units in list A and B can be identified (“Nonresponse”). Neither list A nor B contain overcoverage errors (“Spurious Events”). Finally, under the assumption that the event of being enumerated in list A is independent of that in B (“Causal Independence”), the probability p_{ab} is given by

$$p_{ab} = p_{a+}p_{+b} \tag{1}$$

For application to census undercoverage adjustment, let A be the census data and B the independent coverage-survey data. To avoid additional details, we assume that the coverage survey aims to enumerate the *whole* population at the sampled locations, such as census blocks or postcode areas, so that the missing survey enumerations are not due to sample selection, and the estimation below may be repeated for the target population at *each* sampled location. Because there is a time lag between the two list enumerations in practice, one needs to assume that the target population remains the same (“Closure”).

A large-sample estimator of N and (p_{1+}, p_{+1}) in (1) is given by

$$(\hat{N}, \hat{p}_{1+}, \hat{p}_{+1}) = \left(\frac{N_{1+}N_{+1}}{N_{11}}, \frac{N_{11}}{N_{+1}}, \frac{N_{11}}{N_{1+}} \right)$$

(e.g., Wolter 1986). In particular, $\hat{N}$ is the so-called Dual-System Estimator (DSE). Among others this may be motivated as the method-of-moments estimator (MME) based on the set of moment equations:

$$\begin{cases} E(N_{11}) = Np_{1+}p_{+1} \\ E(N_{1+}) = Np_{1+} \\ E(N_{+1}) = Np_{+1} \\ E(N_{00}) = N - E(N_{1+}) - E(N_{+1}) + E(N_{11}) \end{cases} \tag{2}$$

Note that the last equation is merely a tautology since N_{00} is nonobservable, such that there are in effect only three equations.

3. Model with Additional Overcoverage Errors

3.1. Target-List Universe

Erroneous enumerations in census correspond to reports of nonexistent or out-of-scope cases, such as newborns after the census reference period that are mistakenly recorded in the census. Out-of-scope newborns can equally occur in lists originating from administrative sources, such as when the entry time point of a record is misreported. More often, though, erroneous register enumerations happen because an individual leaves the target population without deregistering. For instance, someone may have moved abroad without notifying their general practitioner and thus becomes an erroneous enumeration in the Patient Register for the census. Likewise, the same individual may fail to notify the election office, and become an erroneous enumeration in the Electoral Register, say, until the next time this person takes part in the general election from abroad.

Generally speaking, therefore, it is unlikely to be the case that overcoverage errors are independent across multiple registers. Moreover, erroneous enumerations may be more extensive in the administrative registers than in the census. For example, the Patient Register enumeration of the population of England and Wales is over four percent higher than the Census 2011 population estimate (ONS 2013). In other words, if unaccounted for, erroneous register enumeration is potentially a source of large bias.

The homogeneity model above is defined for the units in the target population alone. Erroneous list enumeration implies that there are units included in list A or B, or both, which are *not* in the target population U. One needs to extend the reference set to the *target-list universe*, denoted by $U^* = U \cup A \cup B$. Let the probability that a unit in U^* belongs to a particular *target-list domain* be given as below:

			List B		
			in	out	
In U	List A	in	p_{111}	p_{110}	p_{11+}
		out	p_{101}	p_{100}	p_{10+}
			p_{1+1}	p_{1+0}	p_{1++}
			List B		
			in	out	
Out of U	List A	in	p_{011}	p_{010}	p_{01+}
		out	p_{001}	—	p_{001}
			p_{0+1}	p_{010}	

Each unit in U^* is assumed to follow independently ("Autonomous Independence") the multinomial distribution ("Multinomial") with probability p_{uab}, for $u, a, b = 1, 0, +$, except for $(u, a, b) = (0, 0, 0)$ which is *not* part of the target-list universe. Let N_{uab} be the size of the corresponding target-list domain, where $N_{000} \equiv 0$, that is a structural zero. The target population is given by $U = U^*_{1++}$ and its size by $N = N_{1++}$ in this notation. Let N_{uab} be observed for $(u, a, b) = (+, 1, 1)$, $(+, 1, 0)$ or $(+, 0, 1)$, that is the matching of list A and B is errorfree ("Matching"), and let all the list units be identified ("Nonresponse").

Thus, all the assumptions of the homogeneity model are retained, except for the three of "Spurious Events", "Closure" and "Causal Independence". This is of course not to say that the other assumptions are all beyond criticism. But they are not dealt with in this article. In particular, we modify the assumption of "Spurious Events" to exclude all other overoverage errors, such as duplicate reports, but allow for erroneous list enumeration. The "Closure" assumption is no longer necessary, because we now allow for erroneous list enumerations. What remains to be explored are the possibilities of replacing the assumption of "Causal Independence" (1).

3.2. Moment Equations Given Additional Survey Enumeration

The seven parameters of the multinomial distribution are not estimable given only three observed list-domain counts N_{+11}, N_{+10} and N_{+01}. Assume that there exists an additional coverage survey, denoted by S, which (I) has *only* undercoverage error so that all the units in S belong to U, and (II) can be matched to list A and B *without* errors.

The following additional notations seem convenient. Let n_{ab} be the observed number of units in S that belong to the list domain (a, b). Assume that the event of being enumerated in S is independent of the inclusion in the lists, such that

$$\pi_S = P\big(i \in S | i \in U^*_{1ab}\big) = P(i \in S) \tag{3}$$

It follows that $E(n_{ab}) = E(N_{1ab})\pi_S$. Consider two possible decompositions

$$E(N_{1ab}) = E(N)P\big(i \in U^*_{1ab} | i \in U\big) = E(N_{+ab})P\big(i \in U | i \in U^*_{+ab}\big) \tag{4}$$

for $(a, b) \neq (0, 0)$. The first conditional probability that unit $i \in U$ is in the list domain (a, b) will be referred to as the corresponding list *catch rate*, short handed as

$$\pi_{ab} = p_{1ab}/p_{1++}$$

for $a, b = 1, 0, +$. The second conditional probability is given by one minus the conditional probability that a unit in the list domain (a, b) is an erroneous enumeration, for $(a, b) \neq (0, 0)$, to be referred to as the corresponding list *error rate* and short handed as

$$\theta_{ab} = p_{0ab}/p_{+ab} = p_{0ab}/(p_{1ab} + p_{0ab})$$

Given that our interest is to see how the erroneous enumerations can be modelled, it will be useful to observe a set of moment equations, conditional on $\mathbf{x} = (x_{11}, x_{10}, x_{01})$ defined by $x_{ab} = N_{+ab}$, given in terms of the list error rates:

$$\begin{cases} E(n_{11}|\mathbf{x}) = x_{11}(1-\theta_{11})\pi_S \\ E(n_{10}|\mathbf{x}) = x_{10}(1-\theta_{10})\pi_S \\ E(n_{01}|\mathbf{x}) = x_{01}(1-\theta_{01})\pi_S \\ E(n_{00}|\mathbf{x}) = \big(E(N|\mathrm{x}) - x_{11}(1-\theta_{11}) - x_{10}(1-\theta_{10}) - x_{01}(1-\theta_{01})\big)\pi_S \end{cases} \quad (5)$$

Notice that, since the unknown quantity $E(N|\mathbf{x})$ appears only in the last equation, this last equation can only be used to derive an estimate of $E(N|\mathbf{x})$ given the other parameter estimates. There are four parameters in the first three equations of (5). At least one additional assumption is needed from the different models, which can be compared to each other in terms of how they transform the first three equations. The strategy now is to examine systematically the possible log-linear models for, respectively, the target universe U, the target-list universe $U*$ and the *list universe*, denoted by $U_L = A \cup B$.

3.3. A Log-Linear Model of U

The list catch rates are defined for the units in U, conditional on which the N_{1ab}s form a two-way contingency table with fixed total N. The saturated log-linear model is

$$\log \pi_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB}$$

(e.g., Agresti 2013). The largest nonsaturated model is given by

$$\lambda_{ab}^{AB} = 0 \Leftrightarrow \pi_{ab} = \pi_{a+}\pi_{+b} \Leftrightarrow \pi_{11}\pi_{00} = \pi_{10}\pi_{01} \quad (6)$$

that is the event of being enumerated in List A is independent of that in B. Given that $E(n_{ab}|N) = N\pi_{ab}\pi_S$, Model (6) implies

$$E(N_{111}|N) = E(N_{11+}|N)E(N_{1+1}|N)/N$$
$$E(n_{11}|N)E(n_{00}|N) = E(n_{10}|N)E(n_{01}|N)$$

the latter of which can be checked given the n_{ab}s.

As discussed previously, one does not really expect (6) to hold for example between the census and the Patient Register, or between the Patient and the Electoral Registers, and so on. Still, to see the implications of (6) on the list error rates, let $\theta_{1+} = p_{01+}/p_{+1+}$ be the probability that a unit in list A is erroneous and $\theta_{+1} = p_{0+1}/p_{++1}$ that a unit is erroneous in list B. Combining (6) with decompositions like (4), we have

$$\frac{(1-\theta_{11})}{(1-\theta_{1+})(1-\theta_{+1})} = \frac{E(x_{1+})E(x_{+1})}{E(x_{11})E(N)} \quad (7)$$

On account of (7), we refer to (6) as an *incidental* model of the list error mechanism, in the sense that it imposes constraints between the list error rate and the target population size N. For instance, under (6), we have $N = E(N_{11+}|N)E(N_{1+1}|N)/E(N_{111}|N)$.

Since $N_{111} \leq N_{+11} = x_{11}$, and $N_{111} = N_{11+} - N_{110} \geq N_{11+} - N_{+10} = N_{11+} - x_{10}$, and $N_{111} = N_{1+1} - N_{101} \geq N_{1+1} - N_{+01} = N_{1+1} - x_{01}$, we must have

$$\frac{E(N_{11+}|N)E(N_{1+1}|N)}{E(x_{11}|N)} \leq N \leq \min\left(\frac{E(N_{11+}|N)E(N_{1+1}|N)}{E(N_{11+}|N) - E(x_{10}|N)}, \frac{E(N_{11+}|N)E(N_{1+1}|N)}{E(N_{1+1}|N) - E(x_{01}|N)}\right)$$

Now that each list error rate is a conditional probability *within* the list universe, such constraints on the target population size are unwarranted in general.

3.4. Log-Linear Models for Target-List Universe

The saturated log-linear model of p_{uab} of the target-list universe U^* is given by

$$\log p_{uab} = \lambda + \lambda_u^U + \lambda_a^A + \lambda_b^B + \lambda_{ua}^{UA} + \lambda_{ub}^{UB} + \lambda_{ab}^{AB} + \lambda_{uab}^{UAB}$$

Without losing generality, we shall set all the λs to zero except those with all their subscripts equal to one. The structural zero cell, that is, $p_{000} = 0$, can be accommodated by dropping the parameter λ, such that the seven parameters of the saturated model are $(\lambda_1^U, \lambda_1^A, \lambda_1^B, \lambda_{11}^{UA}, \lambda_{11}^{UB}, \lambda_{11}^{AB}, \lambda_{111}^{UAB})$.

The largest nonsaturated hierarchical model is the one with $\lambda_{111}^{UAB} = 0$, denoted by $[UA][UB][AB]$, where

$$\begin{aligned}
p_{100} &= \exp\left(\lambda_1^U\right) \\
p_{010} &= \exp\left(\lambda_1^A\right) \\
p_{110} &= \exp\left(\lambda_1^U + \lambda_1^A + \lambda_{11}^{UA}\right) \\
p_{001} &= \exp\left(\lambda_1^B\right) \\
p_{101} &= \exp\left(\lambda_1^U + \lambda_1^B + \lambda_{11}^{UB}\right) \\
p_{011} &= \exp\left(\lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}\right) \\
p_{111} &= \exp\left(\lambda_1^U + \lambda_1^A + \lambda_1^B + \lambda_{11}^{UA} + \lambda_{11}^{UB} + \lambda_{11}^{AB}\right)
\end{aligned}$$

It follows that

$$\log \frac{p_{011}}{p_{111}} = \log \frac{p_{010}}{p_{110}} + \log \frac{p_{001}}{p_{101}} + \log p_{100}$$

The three log ratios correspond to the log odds of list error in list domain $(1,1)$, $(1,0)$ and $(0,1)$, respectively, denoted by $\operatorname{logit}\theta_{11}$, $\operatorname{logit}\theta_{10}$ and $\operatorname{logit}\theta_{01}$, whereas p_{100} is the proportion of target-population units outside of the list universe. In terms of the list error rates, then, the model amounts to the following assumption

$$\operatorname{logit}\theta_{11} = \operatorname{logit}\theta_{10} + \operatorname{logit}\theta_{01} + (\log E(N_{100}) - \log(N_{+++})) \tag{8}$$

which is an incidental model, just like (6). Since there are no compelling reasons why the conditional probabilities of erroneous enumeration within the list universe must depend on the number of target units *outside* of it, Model (8) cannot be of general use.

It is possible to further reduce the log-linear model. But this would only result in incidental models based on implausible assumptions. For instance, under model $[UA][AB]$

with $\lambda_{11}^{UB} = 0$ in addition, we would have

$$\frac{p_{001}}{p_{101}} = \frac{1}{p_{100}} \quad \text{and} \quad \frac{p_{010}}{p_{110}} = \frac{p_{011}}{p_{111}} = \frac{1}{p_{100}\exp\left(\lambda_{11}^{UA}\right)}$$

3.5. *Log-Linear Models for List Universe*

To separate p_{100} from the list error mechanism, consider now modelling the list universe $U_L = A \cup B$ with the conditional probabilities, for $(a, b) \neq (0, 0)$ and $u = 0, 1$,

$$q_{uab} = p_{uab}/(1 - p_{100})$$

The saturated log-linear model of q_{uab} is given by

$$\log q_{uab} = \lambda + \lambda_u^U + \lambda_a^A + \lambda_b^B + \lambda_{ua}^{UA} + \lambda_{ub}^{UB} + \lambda_{ab}^{AB} + \lambda_{uab}^{UAB}$$

Without losing generality, we shall set all the λs to zero except those with all their subscripts equal to one. There are two structural-zero cells in U_L, namely, $q_{000} = q_{100} = 0$, which can be accommodated by dropping the parameters λ and λ_1^U, such that the six parameters of the saturated model are $\left(\lambda_1^A, \lambda_1^B, \lambda_{11}^{UA}, \lambda_{11}^{UB}, \lambda_{11}^{AB}, \lambda_{111}^{UAB}\right)$.

The largest nonsaturated hierarchical model is the one with $\lambda_{uab}^{UAB} = 0$, where

$$\begin{aligned}
q_{010} &= \exp\left(\lambda_1^A\right) \\
q_{110} &= \exp\left(\lambda_1^A + \lambda_{11}^{UA}\right) \\
q_{001} &= \exp\left(\lambda_1^B\right) \\
q_{101} &= \exp\left(\lambda_1^B + \lambda_{11}^{UB}\right) \\
q_{011} &= \exp\left(\lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}\right) \\
q_{111} &= \exp\left(\lambda_1^A + \lambda_1^B + \lambda_{11}^{UA} + \lambda_{11}^{UB} + \lambda_{11}^{AB}\right)
\end{aligned}$$

In terms of the log odds of erroneous enumeration, that is, logit θ_{11}, logit θ_{10} and logit θ_{01}, this amounts to the following assumption, for $(a, b) \neq (0, 0)$,

$$\text{logit } \theta_{ab} = a\gamma_A + b\gamma_B \Leftrightarrow \text{logit } \theta_{11} = \text{logit } \theta_{10} + \text{logit } \theta_{01} \tag{9}$$

This is a 'standard' null second-order interaction assumption, that is, $\lambda_{uab}^{UAB} = 0$, of the three-way classification of the list units. It is *not* an incidental model. Whether or not plausible for the particular data of concern, it is a model that can *not* be disregarded *a priori*, and it can readily be extended to situations involving more than two lists, where the log-linear model of the extended list universe can be put down similarly.

We note that further reduction of Model (9) would only result in less plausible assumptions. For instance, under model $[UA][AB]$ with $\lambda_{11}^{UB} = 0$ in addition, we have

$$\frac{q_{001}}{q_{101}} = 1 \text{ and } \frac{q_{010}}{q_{110}} = \frac{q_{011}}{q_{111}} = \exp\left(-\lambda_{11}^{UA}\right)$$

that is, the error rate is simply 0.5 for the units in B but not A, and it is the same for all the units in A whether they belong to list B or not, which seems unwarranted in general.

3.6. Two Alternative Log-Linear Models for List Universe

So far (9) is the only model of list erroneous enumeration that (i) does not involve incidental assumptions about the target population size, and (ii) can be extended to include more than two lists. When a list error rate is low, its logit does not differ much from its log. For instance, for a ten percent error rate, we have $\text{logit}\, 0.1 = -2.2$ compared to $\log 0.1 = -2.3$. Replacing logit in (9) with log leads to the following log-linear model

$$\log \theta_{ab} = a\alpha_A + b\alpha_B \Leftrightarrow \log \theta_{11} = \log \theta_{10} + \log \theta_{01} \Leftrightarrow \theta_{11} = \theta_{10}\theta_{01} \qquad (10)$$

for $(a, b) = (1, 1), (1, 0), 0, 1)$, that is, the error rate of the units in both A and B is the product of the error rate of the units in only A (but not B) and that of the units in only B (but not A). That is, for $i \in U_L$,

$$P(i \notin U | i \in A \cap B) = P(i \notin U | i \in A \setminus B) P(i \notin U | i \in B \setminus A)$$

Clearly, every extension of (9) to the situation with more than two lists gives rise to a corresponding model (10), as the two differ only in the choice of the link function. Provided low error rates, the two are expected to yield nearly the same fit to the data. But the difference can become greater if some or all of the error rates are appreciable.

Now, consider the scenario where list A and B have high quality so that both have low erroneous enumerations, that is, both $\theta_{1+} = p_{01+}/p_{+1+}$ and $\theta_{+1} = p_{0+1}/p_{++1}$ are small, and both have high catch rates, so that the list domain $(1, 1)$ is much larger than domain (1, 0) or (0, 1). It then seems natural to expect the error rate to be even lower among the units in both A and B, that is, $\theta_{11} < \theta_{1+}$ and $\theta_{11} < \theta_{+1}$, while the error rates among the units that belong to only one list are comparatively high, that is, $\theta_{10} > \theta_{1+}$ and $\theta_{01} > \theta_{+1}$. It is thus worth considering $\theta_{11} = \theta_{1+}\theta_{+1}$ as an alternative to $\theta_{11} = \theta_{10}\theta_{01}$ above, that is,

$$\log \theta_{11} = \log \theta_{1+} + \log \theta_{+1} \Leftrightarrow \theta_{11} = \theta_{1+}\theta_{+1} \qquad (11)$$

The main difference is that θ_{11} can be much lower under (11) than under (10).

It should be noted that Model (11) does *not* belong to the standard log-linear models for cross classified counts based on the concept of conditional independence. The examination of the possible standard log-linear models above empirically verifies this for the two-list setting. Generically speaking, denote by X, Y and Z any three random events. A conditional independence assumption among them must be of the form

$$P(X \cap Y | Z) = P(X|Z)P(Y|Z)$$

that is, the *conditional joint* probability is the product of the *conditional marginal* probabilities. If we put X as erroneous enumeration for $i \in U_L$, and Y as its inclusion in list A and Z as its inclusion in B, then (11) has the form

$$P(X|Y \cap Z) = P(X|Y)P(X|Z)$$

that is, the *joint conditional* probability is the product of the *marginal conditional* probabilities. We refer to this as an assumption of *pseudoconditional independence* (PCI).

It is possible to develop classes of log-linear models that extend (11) to list CR data involving more than two lists. But we shall not go into the details here. Instead, let us look at a heuristic example of why Model (11) may be more suitable than (10) *when the quality of the list enumerations is high*. Assume two lists that have *no* erroneous enumerations at all and $N_{+11} = N_{+1+} = N_{++1}$, in which case we have $\theta_{11} = \theta_{1+} = \theta_{+1} = 0$ while $(\theta_{10}, \theta_{01})$ do not exist. In other words, Model (11) holds but (10) is not applicable. Suppose now two units leave the population. First, in the ideal case, the two events are registered in both lists so that $(N_{+11}, N_{+1+}, N_{++1})$ are all reduced by two. Then, Model (11) still holds and (10) remains inapplicable. Next, suppose some lack of updating, such that the one event is registered in list A but not B, and the other is registered in B but not A. Then, we still have $\theta_{11} = 0$, but $\theta_{10} = \theta_{10} = 1$, and $\theta_{1+} = 1/(N_{+1+} - 1)$ and $\theta_{+1} = 1/(N_{++1} - 1)$. Model (10) errs much more than (11), because the difference between $\theta_{11} = 0$ and $\theta_{10}\theta_{01} = 1$ is much larger than the difference between $\theta_{11} = 0$ and $\theta_{1+}\theta_{+1} = 1/[(N_{+1+} - 1)(N_{++1} - 1)]$. One can go through the other possibilities of imperfect updating, and one will find that the Model (11) either holds or errs only little.

Both Model (10) and (11) can be fitted given survey data S. For the two-list setting, it is convenient to derive the MME from (5) directly (Appendix). We have

$$\hat{\theta}_{10} = \frac{x_{01}}{n_{01}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{10}}{x_{10}}\right) \quad \text{and} \quad \hat{\theta}_{01} = \frac{x_{10}}{n_{10}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{01}}{x_{01}}\right) \tag{12}$$

for Model (10), and

$$\hat{\theta}_{1+} = \frac{x_{+1}}{n_{+1}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{1+}}{x_{1+}}\right) \quad \text{and} \quad \hat{\theta}_{+1} = \frac{x_{1+}}{n_{1+}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{+1}}{x_{+1}}\right) \tag{13}$$

for Model (11). Any estimated error rate that is negative will be replaced by 0.

4. Simulations

4.1. Range of Fitting

First we explore numerically the differences between the models outlined above, in order to better appreciate the conditions under which a good fit can be achieved for list CR data.

Consider the two-list CR data in Table 1. In Example (I), the number of units is 1,000 in list A and 1,200 in B and 900 in both A and B. The number of erroneous units is 50 in list A and 80 in B. The number of erroneous units among those in both A and B is left to vary, denoted by r_{11}. The number of erroneous units among those in A but not B is then $50 - r_{11}$,

Table 1. Two numerical examples of two-list CR data with under- and overcoverage

		A	B	A and B	A but not B	B but not A
(I)	List enumeration	1,000	1,200	900	100	300
	No. erroneous units	50	80	r_{11}	$50 - r_{11}$	$80 - r_{11}$
		A	B	A and B	A but not B	B but not A
(II)	List enumeration	1,200	1,350	900	300	450
	No. erroneous units	250	400	r_{11}	$250 - r_{11}$	$400 - r_{11}$

Table 2. Values of r_{11} at which models fit perfectly for data in Table 1

	Example (I)		Example (II)	
Model	r_{11}	$(\theta_{10},\ \theta_{01},\ \theta_{11})$	r_{11}	$(\theta_{10},\ \theta_{01},\ \theta_{11})$
(9)	33	(0.170, 0.157, 0.0367)	184	(0.220, 0.480, 0.207)
(10)	30	(0.200, 0.167, 0.0333)	155	(0.317, 0.544, 0.172)
(11)	3	(0.470, 0.257, 0.0033)	56	(0.208, 0.296, 0.062)

and it is $80 - r_{11}$ among those in B but not A. By varying r_{11}, the idea is to see when the Models (9), (10) and (11) appear most plausible. The case is similar for Example (II).

More specifically, for Example (I), Model (9) fits the CR data perfectly when, for some $1 \leq r_{11} \leq 49$, we have $\text{logit}(r_{11}/900) = \text{logit}((50 - r_{11})/100) + \text{logit}((80 - r_{11})/300)$, which occurs at $r_{11} = 33$. Model (10) fits perfectly at $r_{11} = 30$, where $\log\,(r_{11}/900) = \log\,((50 - r_{11})/100) + \log\,((80 - r_{11})/300)$, whereas Model (11) fits perfectly at $r_{11} = 3$, where $\log\,(r_{11}/900) = \log\,(50/900) + \log\,(80/1200)$. The corresponding errors rates are summarized in Table 2. Similarly for Example (II).

The situations that are favorable to Models (9) and (10) are seen to be fairly similar for relatively low error rates such as in Example (I). The one fits best at $r_{11} = 33$ and the other at 30. However, the difference between the two becomes larger as the error rates increase. In Example (II), the one fits best at $r_{11} = 184$ and the other at 155. Also the corresponding error rates are seen to differ more in this case.

Next, Model (11) is more suitable in situations where relatively more erroneous enumerations occur among the units that belong to only one list, while erroneous enumeration is much less probable for units in both lists. In Example (I), the PCI assumption (11) fits best when $r_{11} = 3$ and $\theta_{11} = 0.0033$, the latter of which is *much* lower than the marginal error rates $\theta_{1+} = 0.050$ and $\theta_{+1} = 0.067$. The contrast between θ_{11} on the one hand and $(\theta_{10}, \theta_{01})$ on the other is much larger than under model (9) or (10). The contrast is reduced as the error rates increase in Example (II). But the situation where Model (11) would be plausible is still quite different from those for the other two models.

In conclusion, both Models (10) and (11) are additions to the standard log-linear model (9) rooted in the concept of conditional independence. In particular, Model (11) provides an alternative in situations where there is a large contrast between the overcoverage error among the units in both lists and that among the units in only one list. The aim of the discussion above is to illustrate when the different models might be applicable and how they relate to each other.

4.2. *Adjustment of Census Erroneous Enumeration*

As mentioned earlier, adjustment of census erroneous enumeration traditionally requires a separate O-sample in addition to the independent U-sample for undercoverage adjustment. In theory, an O-sample selected from the list enumerations can be used to estimate the error rates $(\theta_{11}, \theta_{01}, \theta_{10})$. This requires making a strong assumption that fieldwork is able to identify *all* the erroneous list enumerations in the O-sample. It would also imply extra cost, although to some extent this can be controlled by the choice of the O-sample size. On both accounts, it seems of interest if the modelling approach considered in this article can potentially provide useful adjustment of census erroneous enumeration *without* the need for conducting the fieldwork. The possibility is explored here.

Assume three datasets: census, denoted by A, register enumeration processed from administrative sources, denoted by B, and an independent undercoverage survey, denoted by S. Without losing generality, we shall suppose that the survey S attempts to enumerate everyone in the selected areas. This yields the two-list one-survey setting in each surveyed area. The following assumptions and observations are worth noting:

- The census erroneous enumeration rate is expected to be relatively low. We assume that the range of the marginal error rate θ_{1+} of the census (i.e., List A) is reasonably covered by the following set of values: $\theta_{1+} = 0.2\%, 0.5\%, 1\%$.
- The register enumeration can have a higher, even much higher, marginal error rate θ_{+1}. We shall explore the following set of values: $\theta_{+1} = 1\%, 5\%, 10\%, 20\%$.
- Provided independent survey (Equation 3), we have $E(n) = E(N)\pi_S = E(N_{1++})\pi_S$ where n is the total survey enumeration, and $E(n - n_{00}) = E(N_{1++} - N_{100})\pi_S$ where n_{00} is the number of individuals enumerated in S that do not belong to list A nor B. Thus, the *overall* list catch rate can be given by

$$\frac{E(N - N_{100})}{E(N)} = \frac{E(N_{1++} - N_{100})}{E(N_{1++})} = \frac{E(n - n_{00})}{E(n)}$$

  and estimated by $1 - n_{00}/n$, irrespective of the error rates. An important implication is that the *relative* bias induced by the misspecification of a *nonincidental* erroneous enumeration model is unrelated to the target population size N.

- Provided the theoretical value of θ_{11} in addition to θ_{1+} and θ_{+1}, a straightforward simulation approach to evaluate the potential bias of an error model is to repeatedly generate $\mathbf{n} = (n_{11}, n_{10}, n_{01}, n_{00})$ under some given value of π_S, conditional on the target-list universe, and calculate the average of $\hat{N}$ over all the repetitions. More convenient, however, is to fit the moment Equations (5) just *once* to the expected values of $\mathbf{n}$, denoted by $\dot{\mathbf{n}}$, and use the difference between the corresponding $\hat{N}(\dot{\mathbf{n}})$ and N as an approximation to the model bias. This has two advantages: firstly, it makes it clear that the result is invariant to the arbitrary choice of π_S, which cancels out on both sides of the equations in (5) at $\dot{\mathbf{n}} = E(\mathbf{n}|U^*)$; secondly, the result is not subjected to the Monte Carlo errors of the repeated sampling approach.

For comparison to the equally cost-efficient approach without extra fieldwork associated with the O-sample, we consider the DSE based on census A and undercoverage survey S, that is ignoring the potential erroneous census enumerations. Corresponding to the expected survey enumeration $\dot{n}$, this is given by

$$\dot{N}_{DSE} = \dot{n}x_{1+}/\dot{n}_{1+} \approx E(\hat{N}_{DSE}|U^*)$$

Clearly, the relative bias of this unadjusted DSE is simply θ_{1+}, because the hypothetical unbiased DSE is then given by $\dot{n}x_{1+}(1 - \theta_{1+})/\dot{n}_{1+}$.

Table 3. Range of relative bias under Model (10) and (11) for census enumeration error adjustment. Census enumeration = 1,000, register enumeration = 1,200, census-register enumeration = 900. Error rate of census errra (θ_{1+}), register enumeration (θ_{+1}), census-register enumeration (θ_{11}), where 0 < θ_{11} < θ_{1+}. All numbers in %.

Model (10)	Register error rate			
Census error rate	1	5	10	20
0.2	(0.078, 0.078)	(−0.11, −0.11)	(−0.48, −0.48)	(−3.4, −3.4)
0.5	(−0.038, 0.43)	(−0.88, 0.32)	(−2.5, 0.095)	(−16, −1.6)
1	(−0.25, 1)	(−2.3, 1)	(−6.3, 1)	(−38, 1)
Model (11)	**Register error rate**			
Census error rate	1	5	10	20
0.2	(0.11, 0.11)	(0.11, 0.11)	(0.1, 0.1)	(0.089, 0.089)
0.5	(0.11, 0.45)	(0.091, 0.44)	(0.068, 0.44)	(0.014, 0.43)
1	(0.1, 1)	(0.065, 1)	(0.012, 1)	(−0.11, 1)

Table 3 gives the range of relative bias under the Model (10) and (11), respectively. For each combination of (θ_{1+}, θ_{+1}), the number of erroneous enumeration N_{011} among the units in both A and B (i.e., the census-register enumeration) is bounded upwards by min $(N_{+1+}\theta_{1+}, N_{++1}\theta_{+1})$ for the given target-list universe. In the simulation setting here, this is always equal to the integer $N_{+1+}\theta_{1+} = x_{1+}\theta_{1+}$. Each possible N_{011} yields a different target population size $N = N_{1++}$, a corresponding 'joint' error rate $\theta_{11} = N_{011}/x_{11} = N_{011}/N_{+11}$, and a set of expected survey enumerations $\dot{\mathbf{n}}$. The relative bias of a model is given by $\hat{N}(\dot{\mathbf{n}})/N - 1$, where $\hat{N}$ is derived from (12) under Model (10) and (13) under Model (11). As explained above, this relative bias is invariant towards any arbitrary but admissible choice of the survey catch rate π_S and the overall list catch rate adopted in the simulation. The relative biases corresponding to $N_{011} = 1$ and $N_{011} = x_{1+}\theta_{1+} - 1$, respectively, yield the range of relative bias reported in Table 3.

Take first the results for Model (10) in the upper half of Table 3. At $\theta_{1+} = 0.2\%$ and with census enumeration being 1,000, there are only two erroneous census enumerations, and the DSE has a relative bias of 0.2%. Only $N_{011} = 1$ is in the range to be examined, so that the lower and upper ends of the relative bias range coincide in this case. As the register error rate θ_{+1} increases, the estimate of N_{011} increases under Model (10), to the extent that it is 31.6 when the register error rate is 20%, leading to a large negative bias −3.4% due to model misspecification. Next, at $\theta_{1+} = 0.5\%$, the two end points correspond to $N_{011} = 1$ and $N_{011} = 4$. Model (10) is most misleading at the lower end, as the exploration in Subsection 4.1 has indicated, where the estimate of N_{011} is 142.6, leading to a disastrous negative relative bias for N. The performance becomes even worse at $\theta_{1+} = 1\%$, where large negative bias already occurs somewhere between $\theta_{+1} = 1\%$ and 5%. At the upper end, where $N_{011} = 9$, the MME (12) is initially negative and needs to be truncated to 0, that is, no census erroneous enumeration at all. The model estimate $\hat{N}$ then becomes the same as the DSE, and has the same relative bias which is equal to θ_{1+}.

In short, when misspecified, Model (10) can lead to grave negative bias in situations where both the census and the register have non-negligible error rates but the error rate is much lower among the census-register enumeration. For example, at $(\theta_{1+}, \theta_{+1}) = (1\%, 5\%)$, the negative bias of Model (10) would be larger in absolute value than the bias of the DSE for all $\theta_{11} < 0.4\%$.

Turning now to Model (11), we notice immediately that its bias is in *no* case larger than that of the DSE. At $\theta_{1+} = 0.2\%$ and $N_{011} = 1$, the estimate of N_{011} increases from 0.007 at $\theta_{+1} = 1\%$ to 0.2 at $\theta_{+1} = 20\%$. In absolute terms, however, such differences have essentially no bearing on the resulting bias, which is about half of that of the DSE across the range of θ_{+1}. Next, at $\theta_{1+} = 0.5\%$, the model predicted value of N_{011} would be somewhere between 0 and 1 for all the values of θ_{+1} here. As N_{011} increases from 1 and 4, the fitted N_{011} (and N_{01+}) decreases steadily towards 0, resulting in the bias to increase towards that of the DSE. The case is similar at $\theta_{1+} = 1\%$, where Model (11) removes almost all the bias of the DSE as $N_{011} \to 1$, while tending towards the DSE as $N_{011} \to 9$.

Thus, it looks like Model (11) is a more robust choice than (10) for potential adjustment of census erroneous enumeration using an additional list enumeration derived from administrative sources. Within the plausible range of marginal error rates of the census and register enumerations (e.g., in Table 3), the PCI assumption (11) removes essentially all the bias of the census-survey DSE as the number of erroneous enumerations among the units in both the census and the register (i.e., N_{011}) tends to zero. At the other other end, as the latter tends towards its upper bound, that is, $N_{011} \to \min\ (N_{01+}, N_{0+1})$, the bias of the model estimate increases towards that of the DSE.

5. Summary and Discussion

Above we have considered some approaches to modelling erroneous enumeration as a type of overcoverage error. Two types of nonincidental models of the list universe are identified. The first of these consists of standard log-linear models, such as (9), and the associated models using alternative link functions, such as (10). The second of these refers to a class of log-linear models that build on the concept of pseudoconditional independence. The two types of models are suitable for different error mechanisms of the data, and are therefore complementary to each other in practice.

One possible application is the adjustment of census erroneous enumeration based on an independent coverage survey and an additional register enumeration processed from administrative sources. Simulations under what seems to be the plausible range of the census and register error rates suggest that Model (11) is robust towards misspecification of the error rate among the ones enumerated in both the census and the register. The potential bias is bounded upwards by the bias of the DSE that ignores erroneous enumeration.

Of course, further investigation should also take into account the variance of the DSE compared to that of the adjusted model estimator. Simulation on the historic census and register data will be necessary. Moreover, it is important to consider the over and undercoverage adjustments hand in hand. Various authors have considered the so-called triple-system estimator (TSE) based on census, register and coverage survey for adjusting under-coverage. See Griffin (2014) for a recent update. A traditional motivation for the TSE is the possibility to relax the "Causal Independence" assumption (1).

An independent survey, however, is needed in the two-list setting that allows for overcoverage errors. There is simply not enough degree of freedom otherwise. The tension needs to be resolved.

An approach to census-like population statistics without the census is a more ambitious goal. To start with, the census may be replaced by an "improved administrative file" (i.e., register), as some countries have done already. A modelling approach can be used to assess and potentially adjust the erroneous register enumeration, provided very little or no fieldwork associated with the O-sample. It also opens up the possibility for using several input registers instead of one combined register.

Appendix

Method-of-Moment Estimator (MME)

Dividing the first equation in (5) by the second and third, respectively, we obtain

$$\begin{cases} n_{11}(x_1 - r_1) = n_1(x_{11} - r_{11}) = n_1 x_{11}(1 - (r_1 r_2)/(x_1 x_2)) \\ n_{11}(x_2 - r_2) = n_2(x_{11} - r_{11}) = n_2 x_{11}(1 - (r_1 r_2)/(x_1 x_2)) \end{cases}$$

where $(n_1, n_2) = (n_{10}, n_{01})$, $(x_1, x_2) = (x_{10}, x_{01})$ and $(r_1, r_2) = \left(x_{10}\hat{\theta}_{10}, x_{01}\hat{\theta}_{01}\right)$ under Model (10), and $(n_1, n_2) = (n_{1+}, n_{+1})$, $(x_1, x_2) = (x_{1+}, x_{+1})$ and $(r_1, r_2) = \left(x_{1+}\hat{\theta}_{1+}, x_{+1}\hat{\theta}_{+1}\right)$ under Model (11). Note the symmetry between r_1 and r_2. We have

$$ar_1^2 - br_1 + c = 0 \quad \text{where } \left(a, b, c\right) = \left(\frac{n_2}{n_1 x_1 x_2}, \frac{n_{11}}{x_{11} n_1} + \frac{n_2}{n_1 x_2} - \frac{1}{x_1}, \frac{n_{11} x_1}{x_{11} n_1} - 1\right)$$

After some algebra we obtain

$$\Delta = b^2 - 4ac = \left(-\frac{n_{11}}{x_{11} n_1} + \frac{n_2}{n_1 x_2} + \frac{1}{x_1}\right)^2 \quad \text{so that} \quad \frac{b + \sqrt{\Delta}}{2a} \equiv x_1$$

It follows that the admissible r_1 and, by symmetry, r_2 are given by

$$r_1 = \frac{x_2}{n_2}\left(\frac{n_{11}}{x_{11}} x_1 - n_1\right) \quad \text{and} \quad r_2 = \frac{x_1}{n_1}\left(\frac{n_{11}}{x_{11}} x_2 - n_2\right)$$

We obtain r_1/x_1 as $\hat{\theta}_{10}$ under (10) or $\hat{\theta}_{1+}$ under (11). The case is similar for r_2. We obtain $\hat{\theta}_{11}$ according to either Model (10) or (11). Next, we obtain $\hat{\pi}_S = (x_1 - r_1)/n_1 = (x_2 - r_2)/n_2$, and $\hat{N}$ on substituting these parameter estimates into the last equation of (5). Linear approximation yields the variance of the MME.

6. References

Agresti, A. 2013. *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.

Brown, J., O. Abbott, and Paul A. Smith. 2011. "Design of the 2001 and 2011 Census Coverage Surveys for England and Wales." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 174: 881–906. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00697.x.

Cormack, R.M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 409–439. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Griffin, R.A. 2014. "Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020." *Journal of Official Statistics* 30: 177–189. Doi: http://dx.doi.org/10.2478/jos-2014-0012.

Hogan, H. 1993. "The Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060. Doi: http://dx.doi.org/10.1080/01621459.1993.10476374.

IWGDMF – International Working Group for Disease Monitoring and Forecasting. 1995a. "Capture-recapture and Multiple-record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7485050 (accessed 15 July 2015).

IWGDMF – International Working Group for Disease Monitoring and Forecasting. 1995b. "Capture-recapture and Multiple-record Systems Estimation 2: Applications." *American Journal of Epidemiology* 142: 1059–1068. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7485051.

Nirel, R. and H. Glickman. 2009. "Sample Surveys and Censuses." In *Sample Surveys: Design, Methods and Applications*, Vol. 29A, edited by D. Pfeffermann and C.R. Rao. pp. 539–565.

ONS – Office for National Statistics. 2013. *Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice*. ONS Internal Report, available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html (accessed 15 July 2015).

Renaud, A. 2007. "Estimation of the Coverage of the 2000 Census of Population in Switzerland: Methods and Results." *Survey Methodology* 33: 199–210.

Wolter, K. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Received January 2014
Revised November 2014
Accepted November 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 397–414, http://dx.doi.org/10.1515/JOS-2015-0024

Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data

James O. Chipperfield[1] *and Raymond L. Chambers*[2]

Record linkage is the act of bringing together records that are believed to belong to the same unit (e.g., person or business) from two or more files. Record linkage is not an error-free process and can lead to linking a pair of records that do not belong to the same unit. This occurs because linking fields on the files, which ideally would uniquely identify each unit, are often imperfect. There has been an explosion of record linkage applications, particularly involving government agencies and in the field of health, yet there has been little work on making correct inference using such linked files. Naively treating a linked file as if it were linked without errors can lead to biased inferences. This article develops a method of making inferences for cross tabulated variables when record linkage is not an error-free process. In particular, it develops a parametric bootstrap approach to estimation which can accommodate the sophisticated probabilistic record linkage techniques that are widely used in practice (e.g., 1-1 linkage). The article demonstrates the effectiveness of this method in a simulation and in a real application.

Key words: Record linkage; measurement error; parametric bootstrap.

1. Introduction

Record linkage is the act of bringing together records from two or more files that are believed to belong to the same unit in a defined population (e.g., a person or business). Record linkage is an appropriate technique when these data sets are joined to enhance dimensions such as time and breadth or depth of detail. In particular, record linkage is an intrinsic part of virtually all coverage error estimation and correction methodologies, where records from two or more frames, each with incomplete coverage of a target population, are linked in order to estimate the extent of the overlap of these frames. In such cases, coverage error models are usually based on the linked data. Ideally, the linkage will be perfect, that is, all records belonging to the same unit are linked and there are no links between records that belong to different units. However, in many situations perfect linkage is not possible. This is because linking fields (e.g., name, address, postcode) may not uniquely identify a unit, legitimately change over time, be missing or contain errors.

[1] Australian Bureau of Statistics, Methodology Division, P O Box 10, Belconnen, Australian Capital Territory 2616 Australia. Email: james.chipperfield@abs.gov.au
[2] University of Wollongong, National Institute for Applied Statistics Research, Northfields Avenue Wollongong, New South Wales, 2500 Australia. Email: ray@uow.edu.au

Acknowledgments: We would like to thank the referees for their attention to detail and expertise. Their efforts significantly improved this article.

Probabilistic record linkage is a widely used approach to record linkage. In probabilistic record linkage (Fellegi and Sunter 1969) each possible link, called a *record pair*, is given a score based on the likelihood that the two records belong to the same unit. Optimisation algorithms are then used to select which record pairs are declared as links. Probabilistic methods for record linkage are now well established (see Herzog et al. 2007; Winkler 2001; Winkler 2005) and there is a range of computer packages available to implement them. A recent example of probabilistic linkage from the Australian Bureau of Statistics (Zhang and Campbell 2012) is the linkage of person records on its 2006 and 2011 Censuses of Population and Housing to facilitate analysis of how characteristics of cohorts change over time. There are many other examples of probabilistic record linkage from statistical agencies, particularly in the area of health data (see the introduction of Kim and Chambers 2012a).

Naively treating a probabilistically linked file as if it is perfectly linked leads to biased inference. Scheuren and Winkler (1993) and Lahiri and Larsen (2005) (referred to as SW and LL hereafter) propose bias-corrected estimators of coefficients in a linear regression model given data from a probabilistically linked file. Chipperfield et al. (2011) consider the analysis of linked binary variables. Building on Chambers (2009), Kim and Chambers (2012a, 2012b) (referred to as KC hereafter) investigate the analysis of linked data using a more general set of models fitted using estimating equations. Kim and Chambers (2012b) review recent development in inference for regression parameters using linked data.

Linkage models form the key feature of all of the above approaches. The linkage model describes the probability that a record on one file is linked to each of the records on another file. For a linkage model to be useful, it must properly take into account how records were linked. SW and LL do not allow for 1-1 linkage, where every record on one file is linked to a distinct and different record on the other, or for linkage in multiple passes or stages, both of which are commonly used in probabilistic record linkage. In theory, KC allows for 1-1 linkage, but imposes strong constraints on the linkage model in order to do so. KC also requires a clerical sample to estimate the parameters of the linkage model, something which is not always available in practice and which itself can be subject to measurement errors.

This article describes an approach to inference using estimating equations that is based on probabilistically linked data where the linked data file is created under the 1-1 constraint. In fact, the proposed method is valid when the linkage is performed in an arbitrary fashion, as long as the linkage process itself is probabilistic and can be replicated. In particular, we argue that replication is straightforward within the probabilistic record linkage framework of Fellegi and Sunter (1969).

Section 2 introduces the basics of record linkage and the linkage model. It describes a bootstrap approach to fitting the linkage model and compares it with the approach of LL. Section 3 describes how this approach to linkage error modelling can be used to bias correct cross tabulations based on linked data, as well as to make correct inference for binary variables. Section 4 demonstrates through simulation that the proposed approach has good bias and coverage properties. Section 5 considers the performance of the proposed approach for estimating regression coefficients in a real example. Section 6 contains some concluding remarks.

2. Linkage and Linkage Models

This section introduces the basics of record linkage. It also defines the linkage error model, which is an essential ingredient for making correct inference with probabilistically linked data, and proposes a bootstrap approach to estimating it. One crucial aspect of our model is that it distinguishes between the process of linking and whether or not the linking is successful. Here, linking denotes the putting together of records to make up the linked data file, that is, the identification of records that are *believed* to come from the same population unit, while successful linking denotes the event that two records *actually* come from the same population unit.

2.1. *A Framework for the Probability Linking of Two Files of the Same Size*

Consider linking two files, file X and file Y, each containing N records that correspond to the same set of units. Let $i = 1,\ldots, N$ denote an arbitrary indexing of the records in X, and similarly let $k = 1,\ldots, N$ denote an arbitrary indexing of the records in Y. A subset of the set of pairs (i, k) is chosen to define the linked records, and we refer to this subset as the set of *linked* pairs. In addition, let $j = 1,\ldots, N$ denote another indexing of the records on Y such that record i in X and record j in Y is a *correct link* when $i = j$. It is important to note that the j index of a record in Y is unknown, since it by definition requires knowledge of the correctly linked data file.

Suppose that there are L linking fields defined by variables that are common to X and Y. We then define $\mathbf{A}^o = \left(\mathbf{A}^o_{11}, \ldots, \mathbf{A}^o_{ik}, \ldots, \mathbf{A}^o_{NN}\right)$ to be the $L \times N^2$ matrix of *observed agreement patterns* for all record pairs, that is $\left(\mathbf{A}^o_{ik}\right)^T = \left(A^o_{ik1}, A^o_{ikl}, \ldots, A^o_{ikL}\right)$, where $A^o_{ikl} = 0$ or 1 if the linked (i,k) pair disagrees or agrees on the lth linking field, respectively. For example, if *first name*, *last name* and *date-of-birth* are the three linking fields, and if the (i,k)th pair agrees on the first two but not on the third, then $\left(\mathbf{A}^o_{ik}\right)^T = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$. Here '1' indicates agreement and '0' indicates disagreement. There are 2^L possible agreement patterns for a record pair.

Define $\mathbf{A} = (\mathbf{A}_{11}, \ldots, \mathbf{A}_{ij}, \ldots, \mathbf{A}_{NN})$ to be the matrix of *unobserved agreement patterns*, where $\mathbf{A}^T_{ij} = \left(A_{ij1}, A_{ijl}, \ldots, A_{ijL}\right)$, with $A_{ijl} = 0$ or 1 if the (i, j)th record pair disagrees or agrees on the lth linking field, respectively. Although $\mathbf{A}$ is simply a rearrangement of the columns of $\mathbf{A}^o$, this rearrangement is indexed by the unobserved j index. $\mathbf{A}$ is therefore a *latent* variable and can be modelled as the outcome of a random process. A common model for $\mathbf{A}$, and one which we use in this article, is often described by the following set of parameters:

- $M_{iil} = \Pr(A_{iil} = 1 | i = j)$ the probability that the value of the lth linking field for record i in X is the same as the corresponding value for linked record i in Y;
- $U_{ijl} = \Pr(A_{ijl} = 1 | i \neq j)$: the probability that the values of the lth linking field for record i in X and record j in Y are the same, given $i \neq j$.

The probabilities M_{il} and U_{ijl} are often assumed to be *homogeneous*, that is, they do not depend on i and j. In such a situation we denote them by M_l and U_l, respectively. *Conditional independence* is often also assumed. Conditional independence means that for any linked pair, agreement on linking field l is independent of agreement on any other linking field l' for all values $l' \neq l$. This is a strong assumption but, as we will see, it can be a reasonable working assumption.

Put $\boldsymbol{\psi} = (M_1, \ldots, M_L, U_1, \ldots, U_L)$. Prior to probabilistic linking, $\boldsymbol{\psi}$ or its estimate, $\hat{\boldsymbol{\psi}}$, is required. In some cases $\boldsymbol{\psi}$ may be known (e.g., if a unique identifier was available from a previous linkage of X and Y). Computing $\hat{\boldsymbol{\psi}}$ using mixture models has been extensively studied (see Larsen and Rubin 2001, who also consider relaxing the conditional independence assumption).

2.2. *The Linkage Process*

Given a value for $\boldsymbol{\psi}$, and a proposed indexing of Y defined by $k = 1, \ldots, N$, Fellegi and Sunter (1969) suggested calculating a weight for the observed (i,k)th pair of the form $W^o_{ik} = \sum_l w^o_{ikl}$, where

$$\begin{aligned} w^o_{ikl} &= ln(M_l/U_l) && \text{if } A^o_{ikl} = 1 \\ &= ln[(1 - M_l)/(1 - U_l)] && \text{if } A^o_{ikl} = 0. \end{aligned}$$

These authors argue that the larger this pair weight, the more likely that the pair is a correct link. These pair weights are then used in an optimisation algorithm to determine the set of (i,k) pairs that are *declared* as links. An obvious objective function to maximise is $O = \sum_{i,k} W^o_{ik} L_{ik}$, where $L_{ik} = 1$ if the (i,k) pair is linked and $\sum_k L_{ik} = 1$ for all i and k. Often a 1-1 constraint is imposed such that $\sum_k L_{ik} = \sum_i L_{ik} = 1$ for all i and k. Also, in practice a linked pair must have a weight that is greater than a cut-off value, c_o, to be declared a link. The value for c_o can be chosen to ensure that the proportion of links that are correct is acceptably high (see Herzog et al. 2007).

To keep computations to a practical level, records on X and Y are often assigned to *blocks*, where only records within the same block form linked pairs. If there is more than one suitable blocking field, linking can often be performed in multiple passes, where a different set of blocking and linking fields is used in each pass. For example, Chipperfield et al. (2011) consider an example with two passes.

2.3. *The Linkage Model*

The result of a 1-1 linkage process is a generally unknown permutation matrix $\mathbf{P} = [\delta_{ij}]$ with (i,j) element δ_{ij} equal to 1 if record i in X is linked to record j in Y and equal to 0 otherwise. By definition of the i and j indices, diagonal entries of 1 on $\mathbf{P}$ indicate correct links. Let $\mathbf{V}^{(X)}$ denote a matrix of values derived from the information in X. We then put

$$E\left(\mathbf{P}|\mathbf{V}^{(X)}\right) = \mathbf{Q}. \tag{1}$$

We refer to (1) as the *linkage model*. Specifically, the linkage model is given by $\mathbf{Q} = [q_{ij}]$ where $q_{ij} = E\left(\delta_{ij}|\mathbf{V}^{(X)}\right)$ is the probability that record i in X is linked to record j in Y, so q_{ii} is the probability of correctly linking to record i in X, and $\sum_j q_{ij} = \sum_i q_{ij} = 1$. Various authors estimate $\mathbf{Q}$ in different ways.

Let I_{ij} denote the indicator for $i = j$. Chambers (2009) considers an 'exchangeable' linkage model, where

$$q_{ij} = \lambda \mathrm{I}_{ij} + (1 - \lambda)(N - 1)^{-1}(1 - I_{ij}). \tag{2}$$

This model is constrained through a single parameter, λ (or one parameter per block, if a blocking strategy is used). In practice, λ is unknown; Chambers (2009) suggests that in such a situation, it can be estimated using a sample of linked pairs that are reviewed clerically and are assigned, without error, as matches (correct links) or nonmatches (incorrect links). However, note that (2) does not explicitly account for how records are linked (e.g., single pass vs. multiple passes), and so may be inadequate when the method of linking leads to heterogeneous correct (and incorrect) linkage probabilities.

LL on the other hand implicitly assume that ordering Y by the j and k indices leads to the same result, and so put $\mathbf{Q} = [q_{ik}]$, where q_{ik} is estimated by $q_{ik}^{(LL)} = p_{ik}/\sum_{k'} p_{ik'}$, where p_{ik} is the probability that the (i,k) pair is a correct link under the model for $\mathbf{A}$ in Subsection 2.1 (see LL, page 223 for an expression for p_{ik} based on a mixture model). By definition, q_{ik} is then the probability that the (i,k) pair is *linked*. Since the probability of linkage of two records is not the same as the probability that these two records, when linked, are correctly linked, the use of q_{ik} as a proxy for q_{ij} is incorrect in general and, as we show later, can lead to significant bias. Moreover, the estimator $q_{ik}^{(LL)}$ does not factor in all of the complexities of the linking process (e.g., 1-1 linkage), with LL (page 226) noting that "It is not entirely clear how to force one-to-one matches and consider probabilities of matching in which two records in one file have a nonzero probability of matching a record in the second file." Goldstein et al. (2012) make a similar proxy assumption, suggesting the estimator $q_{ik}^{(GS)} = W_{ik}^{o}/\sum_{k'} W_{ik'}^{o}$. In the following section, we define a bootstrap approach to estimating $\mathbf{Q}$ for a linkage process which may include 1-1 assignment.

2.4. A Bootstrap Estimator of Q

We assume that the linking process can be characterised as in Subsection 2.2 and that all linking fields on X and Y are known. We also assume that the conditional independence model (see Subsection 2.1) holds and that an unbiased estimator, denoted by $\hat{\boldsymbol{\psi}}$, of the vector of M and U probabilities defined there is available. Note that if either of these assumptions does not hold, then the bootstrap estimator of $\mathbf{Q}$ defined below may well be biased. In particular, following Winglee et al. (2005), we estimate $\mathbf{Q}$ by bootstrap replication of the linking process. This is accomplished by bootstrapping the unobserved agreement pattern matrix, $\mathbf{A}$, assuming that patterns defined by distinct pairs of population units are independently distributed. That is, for each bootstrap realisation of the linking process, we simulate N^2 realisations $\mathbf{A}_{ij}^{*} = \left(A_{ijl}^{*}\right)$ of $\mathbf{A}_{ij}$ such that

$$A_{ijl}^{*} = \begin{cases} B(\hat{M}_l) & i = j \\ B(\hat{U}_l) & i \neq j \end{cases}$$

where $B(\pi)$ denotes an independent realisation of a Bernoulli random variable with success probability π. Note that since bootstrap replication aims to generate agreement patterns that have a similar distribution to the actual unobserved set of agreement patterns **A**, this assumption of independent Bernoulli realisations is a strong one. Alternative models for **A** can be constructed, using the fact that this matrix defines a network connecting the population units. Further research is required on whether this extra level of sophistication is warranted, however.

Given this bootstrap realisation of **A**, we obtain the corresponding bootstrap realisation of the linkage matrix **P** by repeating the linking process using the bootstrap weights $W^*_{ij} = \sum_l w^*_{ijl}$, where

$$\begin{aligned} w^*_{ijl} &= ln\big(\hat{M}_l/\hat{U}_l\big) && \text{if } A^*_{ijl} = 1 \\ &= ln\big[\big(1-\hat{M}_l\big)\big(1-\hat{U}_l\big)\big] && \text{if } A^*_{ijl} = 0. \end{aligned}$$

We then estimate **Q** by averaging over these bootstrap realisations of **P**. That is, we proceed as follows:

1. Repeat the following steps a total of B times:
   a. Generate $\mathbf{A}(b)$ as the bth independent draw of **A** based on $\hat{\boldsymbol{\psi}}$ and an assumption of independent Bernoulli realisations.
   b. Calculate the linking weights, $W^{(b)}_{ij}$ for all i, j and l, as a function of $\hat{\boldsymbol{\psi}}$ and $\mathbf{A}(b)$.
   c. Link X and Y using the $W^{(b)}_{ij}$using the same algorithm that was used to link the original file. Denote the resulting $N \times N$ permutation matrix of actual links by $\mathbf{P}(b;\hat{\boldsymbol{\psi}})$, that is the columns of **P** are indexed by j and element (i,j) of **P** is equal to 1 if record i in X is linked to record j in Y and 0 otherwise. Note that true links then correspond to record pairs where the (i,i) element of **P** is equal to 1.
2. Estimate **Q** by $\hat{\mathbf{Q}}\big(\hat{\boldsymbol{\psi}}\big) = B^{-1}\sum_b \mathbf{P}(b;\hat{\boldsymbol{\psi}})$.

Note that if X and Y are 1-1 linked, then they must also be linked in this fashion in Step 1(c) above.

3. Estimation of Frequency Tables from Linked Data

3.1. A Bias-Corrected Estimator

Let y be a categorical variable recorded on file Y with categories $y = 1, \ldots, u, \ldots, \mathrm{Y}$, and let x be a categorical variable recorded on file X with categories $x = 1, \ldots, t, \ldots, T$. The values of x and y for the correct links are denoted by x_i and y_i, respectively. Given X we can then define the $N \times T$ incidence matrix $\mathbf{I}^{(X)} = \big[I^{(X)}_{it}\big]$, where $I^{(X)}_{it}$ is the indicator for $x_i = t$. Similarly, given Y we can define the $N \times \mathrm{Y}$ incidence matrix $\mathbf{I}^{(Y)} = \big[I^{(Y)}_{iu}\big]$, where $I^{(Y)}_{iu}$ is the indicator for $y_i = u$. The $T \times \mathrm{Y}$ matrix of frequencies of interest is $\mathbf{N} = \big(\mathbf{I}^{(X)}\big)^T \mathbf{I}^{(Y)}$ where the (t,u) element of **N** is N_{tu}.

Assuming independent and identically distributed population units, the probability distribution for (x,y) is multinomial with parameter $\boldsymbol{\pi} = [\pi_{tu}]$, where π_{tu} is the probability that $(x,y) = (t,u)$. Under the multinomial model, $E\big(\mathbf{I}^{(Y)}\big|\mathbf{I}^{(X)}\big) = \Delta$ where $\Delta = \Delta(\boldsymbol{\pi})$ has (i,u) element,

$$E\big(I^{(Y)}_{iu}\big|I^{(X)}_{it}\big) = \pi_{u|t} I^{(X)}_{it}$$

with $\pi_{u|t} = \pi_{tu}\left(\sum_{u'} \pi_{tu'}\right)^{-1}$. Let y_i^* denote the linked value of y for record i in X. The naive estimator of $\mathbf{N}$, which is based on the assumption that all links are correct, is then

$$\mathbf{N}^* = (\mathbf{I}^{(X)})^T \mathbf{I}^{(Y^*)} = (\mathbf{I}^{(X)})^T \mathbf{P}\mathbf{I}^{(Y)},$$

where $\mathbf{I}^{(Y^*)} = \left[I_{iu}^{(Y^*)}\right]$, and $I_{iu}^{(Y^*)}$ is the indicator for $y_i^* = u$.

A key assumption we now make is that of *conditional independence* of the population distribution of y and the linkage error matrix $\mathbf{P}$ given the values in X. This allows us to write

$$E(\mathbf{P}\mathbf{I}^{(Y)}|\mathbf{I}^{(X)}) = E((\mathbf{P}|\mathbf{I}^{(X)}))E(\mathbf{I}^{(Y)}|\mathbf{I}^{(X)}). \tag{3}$$

Given this assumption, the naive estimator $\mathbf{N}^*$ is a biased predictor of $\mathbf{N}$, since

$$\begin{aligned} E(\mathbf{N}^* - \mathbf{N}|\mathbf{I}^{(X)}) &= (\mathbf{I}^{(X)})^T\left\{E(\mathbf{P}|\mathbf{I}^{(X)})E(\mathbf{I}^{(Y)}|\mathbf{I}^{(X)}) - E(\mathbf{I}^{(Y)}|\mathbf{I}^{(X)})\right\} \\ &= (\mathbf{I}^{(X)})^T(\mathbf{Q} - \mathbf{I}_N)\Delta(\boldsymbol{\pi}) \end{aligned}$$

which is nonzero in general. Here $\mathbf{I}_N$ is the identity matrix of order N. Using $\hat{\mathbf{Q}} = \hat{\mathbf{Q}}\left(\hat{\boldsymbol{\psi}}\right)$ as an estimate of $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\psi})$ leads to the bias-corrected estimator

$$\hat{\mathbf{N}}^* = (\mathbf{I}^{(X)})^T\{\mathbf{I}^{(Y^*)} - (\hat{\mathbf{Q}} - \mathbf{I}_N)\Delta(\hat{\boldsymbol{\pi}})\} \tag{4}$$

where $\hat{\boldsymbol{\pi}}$ is an estimate of $\boldsymbol{\pi}$ defined by

1. Initialising $\hat{\boldsymbol{\pi}}$ by $\hat{\boldsymbol{\pi}}^{(0)}$.
2. Computing $\hat{\mathbf{N}}^{*(h)} = \left[\hat{N}_{ut}^{*(h)}\right] = \hat{\mathbf{N}}^*(\hat{\boldsymbol{\pi}}^{(h)})$.
3. Computing $\hat{\boldsymbol{\pi}}^{(h+1)}$ using $\hat{\pi}_{tu}^{(h+1)} = \max(\hat{N}_{ut}^{*(h)}N^{-1}, 0)$.
4. Iterating between Steps 2 and 3 above until convergence.

In all applications reported later in this article, $\hat{\boldsymbol{\pi}}^{(0)}$ was based on the naive estimate $\mathbf{N}^*$ and this iterative scheme always converged.

3.2. Variance Estimation

Let $\hat{N}_{tu}^*(\hat{\psi})$ denote the (t,u) element of $\hat{\mathbf{N}}^*$ given by (4), and put $\mathbf{I}_t^{(X)} = (I_{tu}^{(X)}, 1 \le u \le Y)$, $\mathbf{I}_u^{(Y)} = \left(I_{tu}^{(Y)}, 1 \le t \le T\right)$ and $\mathbf{I}_u^{(Y^*)} = \left(I_{tu}^{(Y^*)}, 1 \le t \le T\right)$. We can write

$$\begin{aligned} Var\left(\hat{N}_{tu}^*(\hat{\psi})\right) &= Var\left\{E\left(\hat{N}_{tu}^*(\hat{\psi})\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y^*)},\mathbf{I}_u^{(Y)};\psi,\pi\right)\right\} + E\left\{Var\left(\hat{N}_{tu}^*(\hat{\psi})\middle|\mathbf{I}_u^{(Y)},\mathbf{I}_u^{(Y^*)},\mathbf{I}_t^{(X)};\psi,\pi\right)\right\} \\ &= Var\left\{E\left(\hat{N}_{tu}^*(\hat{\psi})\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y^*)},\mathbf{I}_u^{(Y)};\psi,\pi\right)\right\} + V_{tu}^{(\hat{\psi})} \end{aligned}$$

where we identify $V_{tu}^{(\hat{\psi})}$ as the component of variance due to estimation of $\boldsymbol{\psi}$. We then make the large sample approximation

$$Var\left\{E\left(\hat{N}_{tu}^*(\hat{\psi})\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y^*)},\mathbf{I}_u^{(Y)};\psi,\pi\right)\right\} \approx Var\left(\hat{N}_{tu}^*(\psi);\psi,\pi\right).$$

Proceeding along the same lines, and using the conditional independence assumption, we have

$$
\begin{aligned}
&Var\left(\overset{*}{\hat{N}}_{tu}(\boldsymbol{\psi});\boldsymbol{\psi},\boldsymbol{\pi}\right)=Var\left\{E\left(\overset{*}{\hat{N}}_{tu}(\psi)\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y)};\psi\right);\pi\right\}+E\left\{Var\left(\overset{*}{\hat{N}}_{tu}(\psi)\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y)};\psi\right);\pi\right\}\\
&=V_{tu}^{(y)}+E\left\{Var\left(\overset{*}{\hat{N}}_{tu}(\psi)\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y)};\psi\right);\pi\right\}\\
&=V_{tu}^{(y)}+E\left\{Var\left(\overset{*}{\hat{N}}_{tu}(\psi)\middle|\mathbf{I}_t^{(X)};\psi\right);\pi\right\}\\
&=V_{tu}^{(y)}+\overset{*}{V}_{tu}
\end{aligned}
$$

where $V_{tu}^{(y)}$ and $\overset{*}{V}_{tu}$ now denote the components of variance, due to the multinomial model and the linkage errors respectively. That is, we write down the large sample approximation

$$Var\left(\overset{*}{\hat{N}}_{tu}(\hat{\psi})\right)\approx V_{tu}^{(y)}+V_{tu}^{*}+V_{tu}^{(\hat{\psi})}. \tag{5}$$

Note that since $\mathbf{Q}$ is determined by ψ, $V_{tu}^{(\hat{\psi})}$ can also be considered to be the component of variance due to estimating $\mathbf{Q}$.

In order to estimate (5), we start by writing down a large sample approximation to $V_{tu}^{(y)}$ of the form

$$
\begin{aligned}
V_{tu}^{(y)} &= Var\left[E\left\{\overset{*}{\hat{\mathrm{N}}}_{tu}(\psi)\middle|\mathbf{I}_t^{(X)},\mathbf{I}_u^{(Y)};\psi\right\};\pi\right]\\
&= Var\left\{E\left(\left(\mathbf{I}_t^{(X)}\right)^T\mathbf{I}_u^{(Y^*)}-\left[\left(\mathbf{I}_t^{(X)}\right)^T(\mathbf{Q}-\mathbf{I}_N)\Delta(\hat{\pi})\right]_{tu}\middle|\mathbf{I}_t^{(X)};\psi\right);\pi\right\}\\
&\approx Var\left\{E\left(\left(\mathbf{I}_t^{(X)}\right)^T\mathbf{I}_u^{(Y^*)}-\left[\left(\mathbf{I}_t^{(X)}\right)^T(\mathbf{Q}-\mathbf{I}_N)\Delta(\pi)\right]_{tu}\middle|\mathbf{I}_t^{(X)};\psi\right);\pi\right\}\\
&= Var\left(\left(\mathbf{I}_t^{(X)}\right)^T\mathbf{Q}\mathbf{I}_u^{(Y)}\middle|\mathbf{I}_t^{(X)};\pi\right)\\
&= \left(\mathbf{I}_t^{(X)}\right)^T\mathbf{Q}Var\left(\mathbf{I}_u^{(Y)}\middle|\mathbf{I}_t^{(X)};\pi\right)\mathbf{Q}^T\mathbf{I}_t^{(X)}.
\end{aligned}
$$

This suggests the plug-in estimator

$$\hat{V}_{tu}^{(y)} = \left(\mathbf{I}_t^{(X)}\right)^T\hat{\mathbf{Q}}\hat{V}\left(\mathbf{I}_u^{(Y)}\middle|\mathbf{I}_t^{(X)};\hat{\pi}\right)\hat{\mathbf{Q}}^T\mathbf{I}_t^{(X)}$$

where $\hat{V}\left(\mathbf{I}_u^{(Y)}\middle|\mathbf{I}_t^{(X)};\hat{\boldsymbol{\pi}}\right)$ is a diagonal matrix with ith diagonal element $(1-\hat{\pi}_{u|t})\hat{\pi}_{u|t}$.

We estimate V_{tu}^{*} by parametric bootstrapping (Lahiri 2003). That is, given $x_i = t$, we first generate S independent values y_{is}, for $s = 1, \ldots, S$, of y_i by making random draws from the multinomial distribution with parameter $\hat{\boldsymbol{\pi}}_{|t} = \left(\hat{\pi}_{u|t}\right)$, setting $I_{iu}^{(Y)}(s)$ equal to the indicator for $y_{is} = u$. If we write the corresponding simulated true incidence matrix as $\mathbf{I}^{(Y)}(s) = \left[I_{iu}^{(Y)}(s)\right]$, the bth bootstrap value for the simulated linked incidence matrix $\mathbf{I}^{(Y^*)}(s)$ is $\mathbf{I}^{(Y^*)}(b,s) = \mathbf{P}(b)\mathbf{I}^{(Y)}(s)$, where $\boldsymbol{P}(b)$ is the bth bootstrap value of $\mathbf{P}$, obtained using the procedure described in Subsection 2.4. Our estimate $\overset{*}{V}_{tu}$ is then

$$\overset{*}{\hat{V}}_{tu} = S^{-1}\sum_s B^{-1}\sum_b\left(\overset{*}{\hat{N}}_{tu}(b,s)-N_{tu}(s)\right)^2,$$

where $\overset{*}{\hat{N}}_{tu}(b,s)$ is the (t,u) element of $\hat{\mathbf{N}}^*(b,s) = (\mathbf{I}^{(X)})^T\left\{\mathbf{I}^{(Y^*)}(b,s)-(\hat{\mathbf{Q}}-\mathbf{I}_N)\Delta(\hat{\pi})\right\}$ and $N_{tu}(s)$ is the corresponding element of $\hat{\mathbf{N}}^*(s) = (\mathbf{I}^{(X)})^T\{\mathbf{I}^{(Y^*)}(s)-(\hat{\mathbf{Q}}-\mathbf{I}_N)\Delta(\hat{\pi})\}$.

Finally, we use a double bootstrap version of the procedure described in Subsection 2.4 to derive an independent bootstrap estimator for $V_{tu}^{(\hat{\psi})}$. Let $s = 1,\ldots, S$ index the multinomial model-based parametric simulations of y described in the preceding paragraphs. Indexing these new double bootstrap replications by $r = 1,\ldots, R$, we proceed as follows:

1. Generate $\mathbf{A}(r)$ so that $\Pr(A_{iil}(r) = 1) = \hat{M}_l$ and $\Pr\left(A_{ijl}(r) = 1\right) = \hat{U}_l$;
2. Calculate $\hat{\boldsymbol{\psi}}(r)$ (and hence $\hat{M}_l(r)$ and $\hat{U}_l(r)$) from $\mathbf{A}(r)$ in the same way that $\hat{\psi}$ was calculated from $\mathbf{A}$;
3. Calculate $\hat{\mathbf{Q}}(r)$ from $\hat{\psi}(r)$ in the same way that $\hat{\mathbf{Q}}$ was calculated from $\hat{\boldsymbol{\psi}}$. Specifically, this involves for $b = 1,\ldots, B$,
   a. Generating $\mathbf{A}(r,b)$ so that $\Pr(A_{iil}(r,b) = 1) = \hat{M}_l(r)$ and $\Pr(A_{ijl}(r,b) = 1) = \hat{U}_l(r)$;
   b. Generating $\mathbf{P}(r,b)$ as a random realisation of the permutation matrix that characterises probabilistic linkage with agreement patterns $\mathbf{A}(r,b)$ and $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}(\boldsymbol{r})$;
4. Calculate $\hat{\mathbf{Q}}(r) = B^{-1}\sum_b \mathbf{P}(r,b)$ and hence $\hat{\mathbf{N}}^*(r,s) = \left[\hat{N}^*_{tu}(r,s)\right]$, where $\hat{\mathbf{N}}^*(r,s) = (\mathbf{I}^{(X)})^T\{\mathbf{I}^{(Y^*)}(s) - (\hat{\mathbf{Q}}(r) - \mathbf{I}_N)\Delta(\hat{\pi})\}$.

Our bootstrap estimate of $V_{tu}^{(\hat{\psi})}$ is then

$$\hat{V}_{tu}^{(\hat{\psi})} = S^{-1}\sum_s R^{-1}\sum_r\left(\hat{N}^*_{tu}(r,s) - \underset{1\le r'\le R}{av}\left\{\hat{N}^*_{tu}(r',s)\right\}\right)^2$$

where *av* denotes average and $\hat{N}^*_{tu}(r,s)$ is the (t,u) element of $\hat{\mathbf{N}}^*(r,s)$. Choice of values for B, S and R were chosen so that, in simulations, the variability in the estimates of each of the three components of variance (see (5)) were negligible.

The frequentist perspective views $\mathbf{N}$ as a fixed population total and y as a fixed quantity. If all records on files X and Y are linked, then, from a frequentist perspective, $V_{tu}^{(y)} = \mathbf{0}$, and the only sources of variation in $\hat{\mathbf{N}}^*$ are due to incorrect linkage and due to estimating the linkage model, $\mathbf{Q}$.

3.3. *Linking Files of Different Sizes*

Consider the general case where X has N records, Y has M records and there are O linked records. There are also no duplicated records on either X or Y. Previously we considered 1-1 linkage, that is, $O = N = M$. Here we consider the two other important cases, $O < N = M$ and $O = N < M$.

3.3.1. Case 1: $L < O = M$

Linking only a subset of records is common in practice because a cut-off, c_o, for linked pair weights is usually enforced. Without loss of generality, we assume that the first O records in X are linked. The estimator of $\mathbf{Q}$ developed in Subsection 2.4 is no longer appropriate, since it is based on the assumption that all records on X are linked. Here we are interested in estimating the $O \times N$ matrix $\tilde{\mathbf{Q}}$ with ith row $\tilde{\mathbf{Q}}_i = (\tilde{q}_{ij})$, where $\tilde{q}_{ij}$ is the probability that record i in X, $i = 1, \ldots, O$ is linked to record j in Y, $j = 1, \ldots, N$. Some suggested methods of estimating $\tilde{\mathbf{Q}}$ are given below.

1. *Purist* approach. This involves estimating $\mathbf{Q}$ as described in Subsection 2.4, but where the bth replicate is only kept if the set of X records linked in the bth replicate is the same as the set of X records that were originally linked. This could be computationally infeasible if many replicates are discarded.
2. *Bayes' Rule*. This first involves estimating $\mathbf{Q}$ as described in Subsection 2.4. Then, conditioning on the set of L linked records in X and using Bayes' Rule, we get $\tilde{q}_{ij} = q_{ij}\left(\sum_{j=1}^{N} q_{ij}\right)^{-1}$.
3. *Exchangeability*. Here we assume an exchangeable structure (see (1)) for the linkage model, conditional on the set of L linked records in X. Accordingly, it follows that $\tilde{q}_{ii} = \lambda = L^{-1}\sum_{i=1}^{L} q_{ij}\left(\sum_{j} q_{ij}\right)^{-1}$ and $\tilde{q}_{ij} = (1-\lambda)(N-1)^{-1}$ for $i \neq j$, where λ is the average probability of a correct link.

For $\hat{N}^{*}_{tu}$ to remain unbiased when we replace $\mathbf{Q}$ with $\tilde{\boldsymbol{Q}}$ in (1), the conditional distribution of y given x for the O linked records must be the same as for all N records in X. That is, nonlinkage is completely at random. It is possible to relax this assumption to some degree and to improve the accuracy of $\hat{N}^{*}_{tu}$ by assuming that the conditional distributions of y given (x, z) for the linked records and for all records on X are the same. Here z is an auxiliary categorical variable defined on X with categories $1 \leq v \leq V$. This is equivalent to assuming that nonlinkage is at random given z. Let N_{tuv} and N^{*}_{tuv} denote the true and linked counts defined by $x \times y \times z$ cross tabulation. Estimating N_{tuv} given N^{*}_{tuv} is straightforward, since we can simply treat (x,z) as a more detailed version of x. Let $\hat{\mathbf{N}}^{*} = [\hat{N}^{*}_{tuv}]$ denote the bias-corrected estimates (4) defined by this more detailed cross tabulation. Our estimate of N_{tu} then is $\hat{N}^{*}_{tu} = \sum_{v} \hat{N}^{*}_{tuv}$. The more general case, where the nonlinkage is not at random (i.e., there is no available z that can be used to make the linked and unlinked distributions of y, x and z the same), requires further research.

3.3.2. Case 2: $O = N < M$

In this case there are more y records than x records, and all x records are linked. Here $\mathbf{Q} = [\mathbf{Q}_{m}, \mathbf{Q}_{\bar{m}}]$ is $N \times M$, where $\mathbf{Q}_{m}$ is the $N \times N$ linkage model for records on Y with a match, $\mathbf{Q}_{\bar{m}}$ is the $N \times (M-N)$ linkage model for records in Y without a match, and $\mathbf{I}^{(Y)}$ is $M \times Y$. Also, $E(\mathbf{I}^{(Y)}|\mathbf{I}^{(X)})$ is undefined for records on Y that do not have a corresponding record on X. This means that we cannot evaluate the expectation of the naive estimator, $\mathbf{N}^{*}$, and hence correct for its bias. To remedy this, let $E(\boldsymbol{I}^{(Y)}|\boldsymbol{I}^{(X)}) = \tilde{\Delta} = \left(\tilde{\Delta}_{m}^{T}, \tilde{\Delta}_{\bar{m}}^{T}\right)^{T}$, where $\tilde{\Delta}_{m}$ and $\tilde{\Delta}_{\bar{m}}$ are the model expectations for the records with and without a match, respectively. That is, the (j,u) element of $\tilde{\Delta}$ is

$$\tilde{\Delta}_{ju} = \begin{cases} \pi_{u|t} & \text{if } x_j = t \text{ and } j \leq N; \\ \tilde{\mu}_{u} & \text{if } j > N; \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

where $\tilde{\mu}_{u}$ is the mean value of $I^{(Y)}_{iu}$ for the $M-N$ records on Y without a match. From (6) it follows that

$$E(\mathbf{N}^{*}|\mathbf{I}^{(X)}) = (\mathbf{I}^{(X)})^{T}(\mathbf{Q}_{m}\tilde{\Delta}_{m} + \mathbf{Q}_{\bar{m}}\tilde{\Delta}_{\bar{m}}).$$

The bias of the naive estimator of $\mathbf{N}$ is therefore

$$\begin{aligned} E\left(\mathbf{N}^* - \mathbf{N}|\mathbf{I}^{(X)}\right) &= (\mathbf{I}^{(X)})^T\{(\mathbf{Q}_m\tilde{\Delta}_m + \mathbf{Q}_{\bar{m}}\tilde{\Delta}_{\bar{m}}) - (\mathbf{I}_N\tilde{\Delta}_m)\} \\ &= (\mathbf{I}^{(X)})^T\left\{(\mathbf{Q}_m - \mathbf{I}_N)\tilde{\Delta}_m + \mathbf{Q}_{\bar{m}}\tilde{\Delta}_{\bar{m}}\right\} \end{aligned}$$

which suggests the bias-corrected estimator (where a ’hat’ denotes an estimate),

$$\hat{\mathbf{N}}^* = (\mathbf{I}^{(X)})^T\left(\mathbf{I}^{(Y^*)} - \left(\hat{\mathbf{Q}}_m - \mathbf{I}_N\right)\hat{\tilde{\Delta}}_m - \hat{\mathbf{Q}}_{\bar{m}}\hat{\tilde{\Delta}}_{\bar{\mathrm{m}}}\right). \tag{7}$$

Here $\hat{\mathbf{Q}}$ (and hence $\hat{\mathbf{Q}}_m$ and $\hat{\mathbf{Q}}_{\bar{m}}$) as well as $\hat{\tilde{\Delta}}_m$ can be calculated using the bootstrap methods outlined earlier. However, this leaves $\hat{\tilde{\Delta}}_{\bar{m}}$ to be evaluated in (7). It would be reasonable to assume that $\tilde{\Delta}_{\bar{m}}$ is known if M is many times larger than N, in which case it could be estimated from all the records in Y (since records without a match would make up the vast majority of records on this file). Alternatively, if X can be assumed to be a random subsample from Y, then we may write $\hat{\tilde{\mu}}_u = \hat{\pi}_{tu}\left(\sum_{t'}\hat{\pi}_{t'u}\right)^{-1}$, which is the marginal mean of $I_{iu}^{(Y)}$.

Combining the above two cases leads to the general case $O < N < M$. Equation (7) can then be used in place of (4) in the bootstrap algorithm described earlier.

3.4. *Inference for Binary Variables*

Finally, we move away from the estimation of frequencies defined by cross tabulations of linked categorical variables to modelling the distribution of a binary variable. Logistic or log-linear models are commonly used with frequency tables (see Hosmer and Lemeshow 2000).

Define $\mathbf{Z}^T = [\mathbf{z}_1, \ldots, \mathbf{z}_w, \ldots, \mathbf{z}_W]$, where $\mathbf{z}_w$ is a binary vector of length K commonly referred to as the wth covariate pattern. Put $\mathbf{T} = (t_1, \ldots, t_w, \ldots, t_W)^T$ and $\mathbf{R} = (r_1, \ldots, r_w, \ldots, r_W)^T$, where t_w and r_w are the numbers of ‘successful’ and ‘unsuccessful’ cases for the wth covariate pattern.

A model for the number of successful cases when population units are independently distributed is

$$E(t_w) = m_w\mu_w,$$

where $\mu_w = g\left(\mathbf{z}_w^T\boldsymbol{\beta}\right)$, $g()$ is the link function, and $m_w = t_w + r_w$ is the total number of cases. A standard estimate of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, is obtained by solving the score equation

$$\mathbf{H} = \mathbf{Z}^T(\mathbf{T} - diag(\mathbf{M})\boldsymbol{\mu}) = \mathbf{0} \tag{8}$$

where $\mathbf{M} = \mathbf{T} + \mathbf{R}$, $\mu = (\mu_1, \ldots, \mu_w, \ldots, \mu_W)^T$ and $\mu_w = \mu_w(\boldsymbol{\beta})$.

Now consider the case where, again due to linkage error, $\mathbf{T}$ and $\mathbf{R}$ are not available. We can define $\mathbf{N}$ (see Subsection 3.1) so that $\mathbf{N} = [\mathbf{T}, \mathbf{R}]$ is of dimension $W \times 2$. This is the situation discussed in Subsection 3.1 for the case where y is binary and the categories of x correspond to the set of covariate patterns. It follows that we can replace $\mathbf{T}$ and $\mathbf{R}$ in (8) by their estimates $\hat{\mathbf{N}}^* = [\hat{\mathbf{T}}^*, \hat{\mathbf{R}}^*]$, where $\hat{\mathbf{N}}^*$ is given by (4). Note that if the model covariates are all observed on X, then $\mathbf{M} = \mathbf{T} + \mathbf{R}$ and $\hat{\mathbf{M}}^* = \hat{\mathbf{T}}^* + \hat{\mathbf{R}}^*$ are the same. In general, however, this will not be the case. A biased-corrected estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^*$, is therefore obtained by solving the adjusted score equation

$$\mathbf{H}_{adj} = \mathbf{Z}^T(\hat{\mathbf{T}}^* - diag(\hat{\mathbf{M}}^*)\mu) = \mathbf{0}. \tag{9}$$

It is easy to see that if $\mathbf{Q}$ is known, then (9) is an unbiased score equation. Using the same arguments as those underpinning (5), a large sample approximation to the covariance matrix of $\hat{\boldsymbol{\beta}}^*$ can be estimated by

$$\hat{\mathbf{V}}(\hat{\beta}^*) = \hat{\mathbf{V}}^{(y)} + \hat{\mathbf{V}}^* + \hat{\mathbf{V}}^{(\hat{\psi})}.$$

Here $\hat{\mathbf{V}}^{(y)} = (\mathbf{Z}^T\hat{\mathbf{V}}\mathbf{Z})^{-1}$, where $\hat{\mathbf{V}}$ is diagonal with wth element $\hat{\mu}_w(1 - \hat{\mu}_w)$, $\hat{\mu}_w = g\left(\mathbf{z}_w^T\hat{\beta}\right)$,

$$\hat{\mathbf{V}}^* = S^{-1}\sum_s B^{-1}\sum_b \left(\hat{\beta}^*(b,s) - \hat{\beta}^*(s)\hat{\beta}^*(b,s) - \hat{\beta}^*(s)\right)^T,$$

where $\hat{\beta}^*(b,s)$ and $\hat{\beta}^*(s)$ are the solutions to (9) when we replace $\hat{\mathbf{T}}^*$ by $\hat{\mathbf{T}}^*(b,s)$ and $\hat{\mathbf{T}}^*(s)$ respectively, and

$$\hat{\mathbf{V}}^{(\hat{\psi})} = S^{-1}\sum_s R^{-1}\sum_r \left\{\hat{\beta}^*(r,s) - \underset{1\le r'\le R}{av}\left(\hat{\beta}^*(r',s)\right)\right\}\left\{\hat{\beta}^*(r,s) - \underset{1\le r'\le R}{av}\left(\hat{\beta}^*(r',s)\right)\right\}^T,$$

where $\hat{\beta}^*(r,s)$ is the solution to (9) when we replace $\hat{\mathbf{T}}^*$ by $\hat{\mathbf{T}}^*(r,s)$.

4. Simulation Results

4.1. The Simulated Data

We simulated data where files X and Y were each comprised of $N = 1{,}000$ records. The variable x was generated independently for each record such that $x = 1$ with probability 0.75 and $x = 2$ otherwise. The variable y then takes the values 1 or 2 and are generated from the multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_{1|1}, \pi_{2|1}, \pi_{1|2}, \pi_{2|2})$. We consider two possible values for $\boldsymbol{\pi}$, $\boldsymbol{\pi}^{(a)} = (0.7, 0.05, 0.05, 0.2)$ and $\boldsymbol{\pi}^{(b)} = (0.6, 0.1, 0.2, 0.1)$.

We consider the logistic model $\Pr(t_l = 1|\mathbf{z}_l) = 1/\{1 + \exp(-\zeta_l)\}$ and $\zeta_l = z_l^T\left(\beta_0, \beta_1\right)$. Define $\delta(.) = 1$ if the argument is true and $\delta(.) = 0$ otherwise. We fit the model to the generated data (y,x) above where the first covariate pattern is $\mathbf{z}_1 = (1,1)$ when $x = 1$ and $\mathbf{z}_2 = (1,0)$ when $x = 2$, and the number of successes and cases for the kth covariate pattern is $t_k = \sum_i \delta(x_i = k, y_i = 1)$, $m_k = \sum_i \delta(x_i = k)$ respectively for $k = 1{,}2$.

The 1,000 records in files X and Y were allocated to 100 blocks with ten records per block. There were five linking fields. In Scenario 1, the five linking fields had $C_l = 5$, 5, 4, 4, 4 categories for $l = 1,..5$, respectively. In Scenario 2 the five linking fields had $C_l = 7$, 7, 7, 6, 6 categories. The value for each linking field in file X was assigned independently and with equal probability from the set of possible categories.

In Scenario 1 the linking fields were assigned $M_l = 0.8$, 0.6, 0.6, 0.6, 0.6. In Scenario 2, these assignments were $M_l = 0.8$, 0.7, 0.7, 0.6, 0.5. The linking fields in file Y, $F_{li}^{(Y)}$, were generated independently for each i and l according to:

$$\begin{aligned} F_{li}^{(Y)} &= f_{li}^{(X)} \text{ with probability } M_l \\ &= \text{a random and equal probability draw from the set} \{\{1, 2, \ldots, \} - \{f_{li}^{(X)}\}\} \text{ otherwise.} \end{aligned}$$

This meant that $U_l = 1/C_l$.

4.2. Linking Under 1-1 Constraint

Records were linked under the 1-1 constraint. This involved:

1. Sort all record pairs by their weight from highest to lowest;
2. The first record pair in the ordered list is linked;
3. All record pairs containing either of the records linked in Step 2 are removed from the list;
4. Return to Step 2 until all records are linked.

Of note is that the proportion of correct links was 0.74 and 0.91 in Scenarios 1 and 2, respectively.

4.3. Results

The number of replicates used to estimate the linkage model was $B = 300$. The proposed estimator of $\mathbf{Q}$ was unbiased in both scenarios (e.g., the average value of the diagonal was 0.74 in Scenario 1). In contrast, the LL estimator of $\mathbf{Q}$ was significantly biased in both scenarios – the average values of the corresponding diagonals were 0.5 and 0.62 in Scenarios 1 and 2, significantly different from the corresponding true values of 0.74 and 0.91. The conclusion is that the LL method of estimating $\mathbf{Q}$ performs poorly under 1-1 linkage. Consequently, estimates of the regression coefficients on which they are based would be heavily biased with poor coverage.

In order to measure Coverage (for nominal 95% confidence intervals), Bias (as a percentage of the corresponding true value) and Root Mean Squared Error (RMSE), the various approaches to linkage and analysis were applied to 300 independently generated versions of file X and file Y and $S = 10$. When $\boldsymbol{\psi}$ was unknown, the variation in $\hat{\psi}$ was estimated with $R = 10$. These results are summarised in Tables 1 and 2. The main results are:

- Naive inference, which treats the linked file as if it was perfectly linked, can be significantly biased and has poor coverage properties.
- The proposed method has negligible bias and good coverage properties whether or not $\boldsymbol{\psi}$ (and hence $\mathbf{Q}$) was known.
- The accuracy of the proposed estimator is somewhat reduced when $\boldsymbol{\psi}$ is unknown. For example, in Table 1, Scenario 1 where $\boldsymbol{\pi} = \boldsymbol{\pi}^{(a)}$, the RMSE was 0.012 and 0.014 when $\boldsymbol{\psi}$ was known and unknown, respectively.
- The magnitude of the bias in the naive estimator tends to be higher for regression coefficients compared with frequency counts, even though the underlying data are the same. For example, for Scenario 1 and $\boldsymbol{\pi} = \boldsymbol{\pi}^{(b)}$ the bias in the regression coefficients

Table 1. Root Mean Squared Error (RMSE), Relative Bias (RB) and Coverage of nominal 95% confidence intervals generated by different estimators for frequency tables

Linking scenario	$\boldsymbol{\pi}$	Estimator	Is $\mathbf{Q}$ known?	RMSE		RB (%)		Coverage (%)	
				$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{1\|1}$	$\pi_{1\|2}$	$\pi_{1\|1}$	$\pi_{1\|2}$
Perfect linkage	$\boldsymbol{\pi}^{(a)}$	Standard	–	0.009	0.024	0.0	0.1	96	95
1	$\boldsymbol{\pi}^{(a)}$	Standard (naive)	–	0.048	0.14	− 5.1	71.7	1	0
1	$\boldsymbol{\pi}^{(a)}$	Proposed	Yes	0.011	0.030	0.0	0.1	93	96
1	$\boldsymbol{\pi}^{(a)}$	Proposed	No	0.018	0.034	0.0	0.0	92	93
2	$\boldsymbol{\pi}^{(a)}$	Standard (naive)	–	0.013	0.038	2.3	2.4	92	90
2	$\boldsymbol{\pi}^{(a)}$	Proposed	Yes	0.012	0.036	0.1	0.2	94	94
2	$\boldsymbol{\pi}^{(a)}$	Proposed	No	0.014	0.037	0.0	0.0	94	94
Perfect linkage	$\boldsymbol{\pi}^{(b)}$	Standard	–	0.012	0.030	0.0	0.0	95	96
1	$\boldsymbol{\pi}^{(b)}$	Standard (naive)	–	0.036	0.101	5.0	7.2	87	72
1	$\boldsymbol{\pi}^{(b)}$	Proposed	Yes	0.014	0.053	0.0	0.5	94	93
1	$\boldsymbol{\pi}^{(b)}$	Proposed	No	0.017	0.063	0.0	0.0	95	94
2	$\boldsymbol{\pi}^{(b)}$	Standard (naive)	–	0.023	0.057	2.3	8.2	82	66
2	$\boldsymbol{\pi}^{(b)}$	Proposed	Yes	0.016	0.029	0.1	0.1	94	96
2	$\boldsymbol{\pi}^{(b)}$	Proposed	No	0.018	0.037	0	0	93	94

Table 2. Root Mean Squared Error (RMSE), Relative Bias (RB) and Coverage of nominal 95% confidence intervals generated by different estimators for logistic regression

				RMSE		RB(%)		Coverage (%)	
Linking scenario	$\boldsymbol{\pi}$	Estimator	Is **Q** known?	β_0	β_1	β_0	β_1	β_0	β_1
Perfect linkage	$\boldsymbol{\pi}^{(a)}$	Standard	–	0.15	0.20	0.0	0.0	94	97
1	$\boldsymbol{\pi}^{(a)}$	Standard (naive)	–	0.34	0.59	− 22.0	− 13.0	39	22
1	$\boldsymbol{\pi}^{(a)}$	Proposed	Yes	0.17	0.25	0.0	0.0	96	98
2	$\boldsymbol{\pi}^{(a)}$	Standard (naive)	–	0.78	1.30	54.0	34.0	0	0
2	$\boldsymbol{\pi}^{(a)}$	Proposed	Yes	0.20	0.31	1.1	0.0	95	97
Perfect linkage	$\boldsymbol{\pi}^{(b)}$	Standard	–	0.15	0.19	1.7	0.6	95	96
1	$\boldsymbol{\pi}^{(b)}$	Standard (naive)	–	0.78	1.30	− 54.0	− 33.0	0	0
1	$\boldsymbol{\pi}^{(b)}$	Proposed	Yes	0.19	0.29	1.2	0.3	97	97
2	$\boldsymbol{\pi}^{(b)}$	Standard (naive)	–	0.22	0.32	24.0	− 24.0	75	60
2	$\boldsymbol{\pi}^{(b)}$	Proposed	Yes	0.16	0.22	0.1	0.1	96	97

estimates is 54% and -33%, compared with 5% and 7% for frequency count estimates.

- The coverage rates for the naive estimator are sometimes close to the nominal 95% level for estimates of frequency counts, but are consistently lower than 95% for estimates of regression coefficients.

5. Real Example

The 2011 Census of Australian Population and Housing collected economic and social information from over 20 million people living in Australia with a reference date of 9 August 2011 (Census night). The Australian Department of Immigration and Citizenship's (DIAC) collected information about 1,315,000 people who were granted visas to live permanently in Australia from the beginning of 2006 to 9 August 2011; this information is stored on the Settlements Database (SD). Given that the undercoverage of the Census is small (less than 1%) and all migrants in scope of the Census can be identified on the SD, it is reasonable to assume that the records on the SD are a subset of the records on the Census.

Two strategies were used to link the Census and SD (see Richter et al. 2013). The first linking strategy, called *Bronze*, did not use name and address. For the purpose of evaluation we focus on records linked during the fifth linking pass, which used the blocking variables *date of birth* and *sex* and linking variables *country of birth*, *marital status*, *year of arrival in Australia*, *religion*, and *fine-level geography* (see Richter et al. 2013). Probabilistic linking was performed using the 1-1 assignment algorithm in *Febrl* (Christen and Churches 2005). The second linking strategy, called *Gold*, used name and address and required significant evidence in order to assign a link (i.e., high cut off) such that we assume all *Gold* links are correct.

The true proportion of links made by Bronze linkage was $q = 0.64$. That is, 64% of the *Bronze* links were also *Gold* links. Using the replication method in Subsection 2.4, q was estimated to be $\hat{q} = 0.65$. This is a remarkably accurate estimate and suggests that the Fellegi and Sunter (1969) framework, upon which the replication approach was based, is an accurate model for describing linkage errors.

Next we compare the estimator $\hat{N}^{***}$, using the *Bronze* links, to the corresponding population total $\mathbf{N}$, calculated from *Gold* links. This comparison was made using only SD

Table 3. Cross tabulation of proportions according to level of qualification within Visa Class: Based on data obtained by linking settlements database to 2011 Census

		Level of qualification		
Visa class	Estimator	1	2	3
1	True (gold)	0.273	0.642	0.083
	Naive (bronze)	0.220	0.750	0.029
	Proposed (bronze)	0.220	0.755	0.025
2	True (gold)	0.385	0.391	0.222
	Naive (bronze)	0.315	0.641	0.043
	Proposed (bronze)	0.343	0.551	0.105

records that were linked by both *Bronze* and *Gold* – therefore any differences between the two are due only to incorrect links, the error of interest here. After restricting to 30–35-year-olds on the SD there were about 3,000 records. As population estimates are of interest here, y is a fixed quantity and $V_{tu}^{(y)} = \mathbf{0}$.

Table 3 sets out the true and naive frequency tables for *level of qualification* (Census) by *visa class* (SD). For simplicity, frequency counts are expressed as proportions of the marginal counts by *visa class*. Across the Visa Classes, the proposed estimates are closer (measured by the mean absolute difference) to the true proportions when compared with the naive estimates. However, for Visa Class 1 the naive estimates are marginally closer to the true proportions than the proposed estimates. Research into more robust ways of specifying and estimating the parameters in the linkage model is required.

6. Conclusion

Data linkage is being used increasingly by statistical organisations to link administrative data sets. This is because administrative data sets are rich sources of information and linking is a relatively inexpensive process. Probabilistic linking is an approach to linking data sets when there is no unique record key or identifier. This article proposes a method of inference using files that have been probabilistically linked. The method can accommodate 1-1 linking – in fact, as long as the linkage process can be replicated, the proposed method is valid. In this sense, there are good prospects of applying this method to linkage involving multiple passes. The proposed method worked well in a simulation study and showed promise in a real situation.

7. References

Chambers, R. 2009. “Regression Analysis of Probability-Linked Data.” *Statisphere Official Statistics Research Series* 4. Available at: http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm (accessed August, 2013).

Chipperfield, J.O., G. Bishop, and P. Campbell. 2011. “Maximum Likelihood Estimation for Contingency Tables and Logistic Regression With Incorrectly Linked Data.” *Survey Methodology* 37: 13–24.

Christen, P. and T. Churches. 2005. *Febrl – Freely Extensible Biomedical Record Linkage*. Available at: http://cs.anu.edu.au/ ~ Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/contents.html (accessed April 30, 2010).

Fellegi, I.P. and A.B. Sunter. 1969. “A Theory for Record Linkage.” *Journal of the American Statistical Association* 64: 1183–1210. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Goldstein, H., K. Harron, and A Wade. 2012. “The Analysis of Record-Linked Data Using Multiple Imputation With Data Value Priors.” *Statistics in Medicine* 31: 3481–3493. Doi: http://dx.doi.org/10.1002/sim.5508.

Herzog, T.N., F.J. Scheuren, and W.E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer.

Hosmer, D.W. and S. Lemeshow. 2005. *Applied Logistic Regression*, (Second Edition). New York: John Wiley and Sons.

Kim, G. and R. Chambers. 2012a. "Regression Analysis Under Probabilistic Multi-Linkage." *Statistica Neerlandica* 66: 64–79. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00509.x.

Kim, G. and R. Chambers. 2012b. "Regression Analysis Under Incomplete Linkage." *Computational Statistics and Data Analysis* 56: 2756–2770.

Lahiri, P. 2003. "On the Impact of Bootstrap on Survey Sampling and Small Area Estimation." *Statistical Science* 18: 199–210.

Lahiri, P. and M.D. Larsen. 2005. "Regression Analysis With Linked Data." *Journal of the American Statistical Association* 100: 222–230. Doi: http://dx.doi.org/10.1198/016214504000001277.

Larsen, M.D. and D.B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41. Doi: http://dx.doi.org/10.1198/016214501750332956.

Richter, K., G. Saher, and P. Campbell. 2013. *Assessing the Quality of Linking Migrant Settlement Records to 2011 Census Data. Methodology Research Papers*, cat. no. 1351.0.55.043. Canberra: Australian Bureau of Statistics. Available at: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/D7A961FD8534DA6FCA257BC900117FBC/$File/1351055043_aug%202013.pdf (accessed December 1, 2013).

Scheuren, F. and W.E. Winkler. 1993. "Regression Analysis of Data Files that are Computer Matched." *Survey Methodology* 19: 39–58.

Winkler, W.E. 2001. *Record Linkage Software and Methods for Merging Administrative Lists. Statistical Research Report Series*, No. RR2001/03, Bureau of the Census. Available at: https://www.vrdc.cornell.edu/info7470/2007/Readings/rr2001-03.pdf (accessed April 30, 2010).

Winkler, W.E. 2005. *Approximate String Comparator Search Strategies for Very Large Administrative Lists. Statistical Research Report Series*, No. RRS2005/02. Bureau of the Census. Available at: https://www.census.gov/srd/papers/pdf/rrs2005-02.pdf (accessed April 30, 2010).

Winglee, M., R. Valliant, and F. Scheuren. 2005. "A Case Study in Record Linkage." *Survey Methodology* 31: 3–11.

Zhang, G. and P. Campbell. 2012. "Data Survey: Developing the Statistical Longitudinal Census Dataset and Identifying Its Potential Uses." *Australian Economic Review* 45: 125–133. Doi: http://dx.doi.org/10.1111/j.1467-8462.2011.00673.x.

Received January 2014
Revised November 2014
Accepted December 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 415–429, http://dx.doi.org/10.1515/JOS-2015-0025

Coverage Evaluation on Probabilistically Linked Data

Loredana Di Consiglio[1] *and Tiziana Tuoto*[1]

The Capture-recapture method is a well-known solution for evaluating the unknown size of a population. Administrative data represent sources of independent counts of a population and can be jointly exploited for applying the capture-recapture method. Of course, administrative sources are affected by over- or undercoverage when considered separately. The standard Petersen approach is based on strong assumptions, including perfect record linkage between lists. In reality, record linkage results can be affected by errors. A simple method for achieving *linkage error-unbiased* population total estimates is proposed in Ding and Fienberg (1994). In this article, an extension of the Ding and Fienberg model by relaxing their conditions is proposed. The procedures are illustrated for estimating the total number of road casualties, on the basis of a probabilistic record linkage between two administrative data sources. Moreover, a simulation study is developed, providing evidence that the adjusted estimator always performs better than the Petersen estimator.

Key words: Linkage errors; capture-recapture method; Petersen estimator; administrative data.

1. Introduction

The problem of assessing the unknown size of a population is one that has long been grappled with, from the first experiments at measuring wild animal population size during the seventeenth century (Petersen 1896; Lincoln 1930) to applications for determining the number of people affected by specific diseases or using illegal drugs (Bartolucci and Forcina 2006), including the population census coverage (Wolter 1986). One well-known and widespread solution for this problem is the capture-recapture method. This method consists of comparing two (or more) independent counts ("capture" in the field of wild animal population estimation) of the same units, then evaluating, without error, the number of individuals in both the counts, and, as a result, counting the number of those caught only once.

In this framework, the standard Petersen estimator works well under some strong assumptions, such as the independence of the lists, the homogeneity of capture probabilities, and the lists' error-free linkage at record level.

Several extensions and adjustments of the Petersen estimator have been proposed over time in order to avoid bias due to failure of these assumptions, which causes the population to be under- or overestimated (e.g., Chao 2001, Chen and Kuo 2001).

Nowadays, the use of administrative data is emerging as a new opportunity in several statistical fields. Administrative data represent sources of several independent counts of a population. They can be exploited for the application of the capture-recapture method to estimate the unknown size of the population.

[1] Italian National Statistical Institute - Istat, Via Cesare Balbo, 16 00184 Rome, Italy. Email: diconsig@istat.it and tuoto@istat.it

Since records are collected for different purposes by different actors, the different administrative sources can be expected to be independent recaptures of the same (sub)population, in contrast to survey data, which are collected by the same organization. In fact, the independence assumption could be violated if the heterogeneity of capture probabilities of units is not properly encompassed in the statistical model.

Given their large size, data sets collected by administrative sources require a massive use of automatic tools, implementing record linkage techniques. Therefore, the error-free linkage assumption can be compromised, particularly in absence of unique identifiers for privacy issues.

In this article, we concentrate on failure of the perfect linkage hypothesis and we analyse different proposals that adjust the Petersen estimator by explicitly taking into account linkage errors.

In Ding and Fienberg (1994), a simple method to achieve linkage error-unbiased estimators of population total and undercoverage rate is proposed; moreover, different models for the two types of linking errors are described. The Ding and Fienberg (1994) adjustment considers the probability of missed true links and the probability of erroneous links, providing an alternative formula with respect to the Petersen estimator to assess the undercoverage and consequently the true population total.

We enhance the Ding and Fienberg (1994) model by defining the probabilities of being counted in both lists, handling the two lists in a symmetric way. These findings are subject to conditions of admissibility, which are discussed in the Appendix. 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: the "Causes of death" register and the "road accidents resulting in deaths (within 30 days) or injuries" register. Simulated data are used to show the benefit of the proposed new method over the existing ones in different linkage scenarios.

2. Capture-Recapture Background

The Petersen model (see Wolter 1986) is a standard well-known model for evaluating the population total. Let N be the unknown population total, and $N_1$and N_2 the population size reported in the first and second list, respectively. Let x_{11} be the number of units recorded in both lists, $x_{12} = N_1 - x_{11}$ the number of units reported only in List 1 and $x_{21} = N_2 - x_{11}$ the number of units reported only in List 2.

The counts can be organised in a 2 X 2 contingency table, with x_{22} the unknown number of units missed by both lists (Table 1).

Under the assumption of independent captures, the number of individuals in the contingency table follows the multinomial distribution.

Table 1. Contingency table of the counts in the two lists

		List 2	
		Present	*Absent*
List 1	*Present*	x_{11}	x_{12}
	Absent	x_{21}	x_{22}

Moreover, adding the following assumptions:

1. the population is closed, so the population being measured in both sources is the same
2. records from both sources can be linked without errors
3. units have the same capture probabilities within each source (homogeneity probability assumption)
4. overcount in both sources is negligible

an unbiased estimator of N, the well-known Petersen estimator, is given by

$$\widetilde{N}_P = N_1 \times N_2 / x_{11}. \tag{1}$$

The first list coverage is then given by

$$\widetilde{\tau}_{1,P} = x_{11} / N_2 \tag{2}$$

and similarly the second list coverage is

$$\widetilde{\tau}_{2,P} = x_{11} / N_1. \tag{3}$$

The previous assumptions' validity has been widely debated in a traditional survey context. Several extensions and adjustments have been proposed in order to avoid biases due to any failure of these assumptions that is under- or overestimation of the real population total amount.

As discussed above, on one hand, the independence of administrative sources could be guaranteed by different data collectors, while on the other hand, the heterogeneity of capture probabilities is a common issue in different settings due to inherent individual behaviour. When the individual capture propensity is not properly modelled, the dependence between lists can arise even in an administrative data context. Much literature focuses on including sources' dependencies and captures' heterogeneity by means of:

- extensions of the log-linear model (Fienberg 1972; Cormack 1989; Chao 2001, Agresti 1994; Coull and Agresti 1999)
- the conditional multinomial logit model (McFadden 1974; Bock 1975; Chen and Kuo 2001; Zwane and van der Heijden 2005)
- the latent class model (Bartolucci and Forcina 2006)
- the Bayesian capture-recapture model (Ghosh and Norris 2005).

More specifically, log-linear models explain the dependencies between data collections and the heterogeneity of capture probabilities by using categorical covariates, while the conditional multinomial logit model also allows continuous covariates to be included in the models.

The latent class model can be considered a conditional multinomial logit model extension and permits the modelling of both the observed heterogeneity using covariates and the unobserved heterogeneity by assuming that units belong to distinct latent classes. Finally, Bayesian capture-recapture models allow dependencies and heterogeneity to be formalised by means of suitable parameters for the distribution of individual capture probabilities.

When dealing with administrative data, compared to the survey context a change of perspective regarding the validity of previous assumptions is needed. In fact, overcoverage in the administrative lists may assume a relevant role. Recently, we have seen the failure of the last assumption 4), due to an observed significant level of list overcoverage affecting administrative data. Large et al. (2011) propose an adjustment to the Petersen estimator in order to correct bias due to overcount within the census context.

Another matter emerging when dealing with administrative sources concerns the unavailability of unique identifiers for maintaining privacy. In this framework, linkage errors could arise. This article considers extensions to deal with record linkage between lists affected by errors.

3. Including Linkage Errors in the Petersen Estimator

In this section, a short description of the most common probabilistic record-linkage framework is given, mainly in order to formalise linkage errors. Moreover, the Ding and Fienberg (1994) estimator to adjust the Petersen one for linkage errors is briefly reported; an extension is introduced to deal with more generic contexts, including those contexts typical for administrative data.

3.1. Linkage Model and Error Evaluation

A key step in applying the Petersen model is the integration of two (or more) sources at record level to identify the common units: this action is commonly referred to as record linkage.

A fundamental theory for record linkage is given in the seminal paper by Fellegi and Sunter (1969). Given two lists, say L1 and L2, of size N_1 and N_2, let $\Omega = \{(\mathrm{a, b}), \mathrm{a} \in \mathrm{L1}$ and $\mathrm{b} \in \mathrm{L2}\}$ be the complete set of all possible pairs, of size $|\Omega| = N_1 \times N_2$. The linkage process between L1 and L2 can be viewed as a classification problem where the pairs in Ω have to be assigned to two independent and mutually exclusive subsets M and U, such that:

M is the link set (a = b)
U is the nonlink set (a ≠ b).

In order to assign the pairs to the sets M or U, K common identifiers (the linking variables) are chosen and, for each pair, a comparison function is applied in order to obtain a comparison vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, . . ., \gamma_K\}$. The ratio r of the conditional probability of $\boldsymbol{\gamma}$ given that the pair belongs to set M to the conditional probability of $\boldsymbol{\gamma}$ given that the pair belongs to set U

$$r = \frac{P(\gamma|(a,b) \in M)}{P(\gamma|(a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)}$$

is used to classify the pairs. The probabilities m and u can be estimated by assuming the true link status is a latent variable, using, for instance, the EM algorithm (Jaro 1989). Hence, those pairs for which r is greater than the upper threshold value T_m are assigned to the set of linked pairs, M^*; those pairs for which r is smaller than the lower threshold value T_u are assigned to the set of unlinked pairs U^*; if r falls in the range (T_u, T_m), no decision is made and the pair is checked by clerical review.

The thresholds are chosen to minimise false link probability, β, and false nonlink probability, $1-\alpha$, defined as follows:

$$\beta=\sum_{\gamma\in\Gamma}u(\gamma)P(M^*|\gamma)=\sum_{\gamma\in\Gamma_{M^*}}u(\gamma)\quad\text{where}\quad\Gamma_{M^*}=\{\gamma:T_m\leq m(\gamma)/u(\gamma)\}\tag{4}$$

$$1-\alpha=\sum_{\gamma\in\Gamma}m(\gamma)P(U^*|\gamma)=\sum_{\gamma\in\Gamma_{U^*}}m(\gamma)\quad\text{where}\quad\Gamma_{U^*}=\{\gamma:T_u\geq m(\gamma)/u(\gamma)\}.\tag{5}$$

The linkage model also provides an evaluation of the probability of a link being a correct given that the link is assigned, the so-called true match rate:

$$\eta=1-\frac{\sum\limits_{\gamma\in\Gamma_M}u(\gamma)P(M^*|\gamma)}{\sum\limits_{\gamma\in\Gamma_M}m(\gamma)P(M^*|\gamma)}=1-\frac{\sum\limits_{\gamma\in\Gamma_{M^*}}u(\gamma)}{\sum\limits_{\gamma\in\Gamma_{M^*}}m(\gamma)}.\tag{6}$$

3.2. *The Ding and Fienberg Estimator*

In the context of probabilistically linked data, the coverage rates and population total estimates produced by the Petersen model may be biased and so they need to be "adjusted" in order to explicitly take into account the linkage errors.

A simple method for achieving "linkage error-unbiased" estimators of the population total and the coverage rates has been suggested by Ding and Fienberg (1994). They relax the perfect linkage assumption, propose models to describe linking errors and include those errors in the estimators derived by the Petersen model.

Under the following assumptions:

(a) true links between L1 and L2 are assigned with probability α
(b) false links between records belonging to M (see Subsection 3.1) are negligible
(c) false links can occur with a common probability β between records belonging to U (see Subsection 3.1)
(d) linkage direction from L1 to L2,

the adjustment proposed by Ding and Fienberg (1994) considers the false nonlink of linking cases probability (i.e., the probability of missing true link, 1-α) and the false link of nonlinking case probability (i.e., the probability of linking false pairs, β),

$$\widetilde{N}_{DF}=\frac{N_{1\cup2}}{\hat{\tau}_{1,DF}+\hat{\tau}_{2,DF}-(\alpha-\beta)\hat{\tau}_{1,DF}\hat{\tau}_{2,DF}-\beta\hat{\tau}_{1,DF}}\tag{7}$$

where $\hat{\tau}_{1,DF}$ and $\hat{\tau}_{2,DF}$ are the estimates of probabilities of being recorded in lists 1 and 2, respectively. $N_{1\cup2}=x_{11}+x_{12}+x_{21}=x_{11}^*+x_{12}^*+x_{21}^*$ is the number of records in list 1 or list 2, with x_{11} the number of *true* records in both lists, x_{12} the number of *true* records in list 1 and not in list 2 and, vice versa, x_{21} the number of *true* records in list 2 and not in list 1, while $x_{11}^*, x_{12}^*, x_{21}^*$ are the observed number of records in both lists, in list 1 and not in list 2, and in list 2 and not list 1, respectively, resulting from the linkage procedure.

The coverage of the first list is given by:

$$\hat{\tau}_{1,DF} = \frac{-x^*_{11} + \beta(x^*_{11} + x^*_{12})}{(\beta - \alpha)\,(x^*_{11} + x^*_{21})} \tag{8}$$

and similarly the coverage of the second list is

$$\hat{\tau}_{2,DF} = \frac{-x^*_{11} + \beta(x^*_{11} + x^*_{12})}{(\beta - \alpha)(x^*_{11} + x^*_{12})}. \tag{9}$$

The coverage rate estimates, $\hat{\tau}_{1,\mathrm{DF}}$ and $\hat{\tau}_{2,DF}$, are obtained by maximizing the conditional likelihood of $(x^*_{11}, x^*_{12}, x^*_{21})$ given $N_{1\cup 2}$,

$$L_1(p_{11}, p_{12}, p_{21}) = L_1(\tau_1, \tau_2) = \frac{N_{1\cup 2}!}{x^*_{11}!x^*_{12}!x^*_{21}!}\frac{p_{11}^{x^*_{11}}p_{12}^{x^*_{12}}p_{21}^{x^*_{21}}}{(p_{11} + p_{12} + p_{21})^{N_{1\cup 2}}}. \tag{10}$$

In this setting, a record is counted in both lists when it is actually in both lists and a link is made, and when the record is only in L1 but it is incorrectly linked with a record in L2. The former event has the probability $\alpha\tau_1\tau_2$, whereas the latter has $\beta\tau_1(1 - \tau_2)$, so the probability of observing a count in (1,1) is $p_{11} = \alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2)$. The probability of occurrence in cell (1,2) and (2,1) can be derived as $p_{12} = \tau_1 - p_{11}$and $p_{21} = \tau_2 - p_{11}$, respectively. See Ding and Fienberg (1994) for more details.

Note that the solutions are admissible under conditions on relationships of errors and counts.

The previous estimators are based on the assumptions: false links that occur when at least two errors are made (that is, records are incorrectly linked and the correct link is missed) have negligible probability of occurrence (assumption b). Moreover, a direction from L1 to L2 is assumed both in the linkage procedure (assumption d) and in the specification of the linkage errors. In the next subsection, generalised estimators for (7)–(9) achieved by relaxing assumption d are illustrated.

3.3. A Generalised Estimator

The Ding and Fienberg (1994) proposal was explicitly defined in the traditional census coverage evaluation context, where the linkage procedure between census data and the postenumeration survey results (Wolter 1986) works in one direction. When dealing with administrative data sources, this assumed one-way linkage direction is not guaranteed. Linkage errors, in particular false links, can occur in both directions, in contrast to what is assumed in d) of Subsection 3.2 according to Model B proposed by Ding and Fienberg (1994, 150). Note that in the context of administrative data, due to differences in unit and time reference, as well as variables' definitions, joint linkage errors (i.e., incorrect link and missed true links at the same time) may occur. Nevertheless, their probability can still be assumed negligible as at least three errors should be made, each one with small probability.

In the present proposal, assumption d) in Subsection 3.2 is relaxed, allowing for two-directional linkage. Hence, the probability of an occurrence in cell (1,1) is $p_{11} = \alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2) + \beta\tau_2(1 - \tau_1)$ where $\alpha\tau_1\tau_2$ is the probability that a unit is actually in both lists and a link is made, $\beta\tau_1(1 - \tau_2)$ is the probability that a unit actually registered only in L1 is incorrectly linked with a record in L2, and finally $\beta\tau_2(1 - \tau_1)$ is the probability that a unit actually registered only in L2 is incorrectly linked with a record in L1. The probability of occurrence in cell (1,2) and (2,1) can be derived as $p_{12} = \tau_1 - p_{11}$ and $p_{21} = \tau_2 - p_{11}$, respectively.

Replacing p_{11}, p_{12}, p_{21} as defined above in the conditional likelihood (10) and maximizing with respect to τ_1 and τ_2, the Modified Ding and Fienberg (MDF) estimators are given by

$$\hat{\tau}_{1,MDF} = \frac{2\beta x^*_{11} + \beta x^*_{12} + \beta x^*_{21} - x^*_{11}}{(2\beta - \alpha)\left(x^*_{11} + x^*_{21}\right)} \tag{11}$$

$$\hat{\tau}_{2,MDF} = \frac{2\beta x^*_{11} + \beta x^*_{12} + \beta x^*_{21} - x^*_{11}}{(2\beta - \alpha)\left(x^*_{11} + x^*_{12}\right)}. \tag{12}$$

Once $\hat{\tau}_{1,MDF}$ and $\hat{\tau}_{2,MDF}$ are obtained, the MDF ML estimator of N is given by:

$$\widetilde{N}_{MDF} = \frac{N_{1\cup 2}}{\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - (\alpha \hat{\tau}_{1,MDF}\hat{\tau}_{2,MDF} + \beta(\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - 2\hat{\tau}_{1,MDF}\hat{\tau}_{2,MDF}))} \tag{13}$$

Conditions for the admissibility of the estimates (11)–(12) also apply (see the Appendix).

The proposed estimators as well as the DF estimators are based on the assumption that linkage errors are constant. If this assumption holds at least in subgroups, the estimators can be applied within strata in which matching error probabilities (and capture probabilities) can be assumed to be more homogeneous than in the whole population.

4. Applications

4.1. Real Data Application

In this section, we present an application to data coming from two independent registers of deaths caused by road accidents. These data are exploited mainly because a complete analysis of the linkage status by clerical review is possible thanks to their small size.

In Italy, police authorities locally collect the road accidents resulting in deaths (within 30 days) or injuries and provide those data to the National Institute of Statistics. The Road Accident Register (denoted as RAR – or list 1, in the following) is an exhaustive, monthly-based register reporting the dynamics and circumstances of road accidents. Data collected by police are the main source for studying road traffic injuries. However, although the police usually collect very detailed information on crash dynamics and circumstances, relevant underreporting could occur due to the very complex situations related to the seriousness of the accidents. Therefore, the integration with health-care databases, such as mortality registers, can be very useful, complementing police data by capturing missing cases and also enriching them with detailed information on causes of death. For this purpose, a record linkage between the RAR and the data on causes of mortality, collected by the Italian National Vital Statistics Death Registry on causes of death (RCD – or list 2, in the following), was carried out.

The linkage procedure is not straightforward: a common personal identifying code is not available. Moreover, since RAR reference units are the road accidents, personal identifying variables (i.e., names, surnames, ages) are sometimes missing or mistaken when more than one person is involved.

The reference year of the application is 2009. As far as the data from RAR are concerned, only records with at least one fatal casualty are considered, corresponding in that year to 4,237 records. Regarding RCD data, only road-accident deaths are considered, according to ICD-10 codes for traffic accidents involving motor vehicles on public roads. These correspond to a total of 4,642 records. The variables used for the linkage are: the road traffic victim/dead person name, surname and age, and the accident/death day, month, municipality and province.

The selected data sources' sizes do not require reduction procedures and the cross product of all records can be considered. The whole linking space is also exploited for the clerical review of links missed by the probabilistic procedure.

The linkage procedure identifies 3,129 linked records. The linkage errors estimated by the Fellegi-Sunter model (see (4) and (5)) are $\beta = 0.00$ and $1 - \alpha = 0.15$.

As is well known, in this approach the accuracy of linkage-error estimates is heavily dependent on the estimates' accuracy in the tails of the $m(\gamma)$ and $u(\gamma)$ distributions. Misspecifications in the model assumptions, errors or lack of information can cause a loss of accuracy in the latter. So, even though in most practical cases the linkage procedure is robust with respect to the links identification, the linkage error-estimates based on the linkage model are nevertheless generally too optimistic (Larsen and Rubin 2001).

As stated above, with these data, a clerical review of the linkage status is possible: this allows an evaluation of the proposed estimators knowing the true linkage-error values.

According to Table 2, the true $1 - \alpha$ is 0.1141 and β is 0.0009. On the basis of the true linkage status, the Petersen estimate of the total amount of road deaths is 5,572.

The results for the population size and the coverage list rates evaluation using the illustrated estimators are summarised in Table 3, where DF and MDF are defined in (7)–(9) and (11)–(13), respectively, and the naïve Petersen estimators are given by Equations (1)–(3), replacing the unobserved count x_{11} by the observed one $\overset{*}{x}_{11}$.

As expected, the DF and the MDF estimators give the same results when linkage errors are obtained from the linkage model due to the negligible value of β. All the compared estimators provide values close to the true one when linkage errors are known. Moreover, they are also less biased than the naïve Petersen estimates when linkage errors are estimated via the linkage model.

It is worth noting that even when a training set with known linkage status is available, the evaluation of β and $1 - \alpha$ is not straightforward. For instance, the well-known method

Table 2. *Comparison between true linkage status and probabilistic linkage results*

		True linkage status		
		Link	Nonlink	
Probabilistic linkage	Link	3,127	2	3,129
	Nonlink	403	2,218	2,621
		3,530	2,220	

Table 3. Amount of road deaths and coverage-rate estimates with estimated and true linkage errors

	True values	Petersen		DF	MDF
N	5,572	6,286	Estimated linkage errors	5,330	5,330
		–	True linkage errors	5,569	5,571
Coverage rate List 1	0.760	0.674	Estimated linkage errors	0.795	0.795
		–	True linkage errors	0.761	0.761
Coverage rate List 2	0.833	0.738	Estimated linkage errors	0.871	0.871
		–	True linkage errors	0.833	0.833

proposed by Belin and Rubin (1995) only provides estimates for β. In fact, detecting false links is more practicable than identifying missing links.

4.2. Simulation Study

The previous section showed an interesting real capture-recapture application that takes into account linkage errors. In that case, even with low linkage-error levels, the adjusted estimators perform better than the naïve Petersen estimator. However, the benefit of the proposed MDF over the DF is not sufficiently evident. In this section, a simulation is performed on fictitious data in order to compare the estimators in different linkage scenarios with variables of varying identifying power.

4.2.1. Description of the Simulated Setting

The simulation study was conducted on 100 replicated settings. Each one consists of a population of 1,000 units and two different lists that are generated mimicking the register undercoverage and the presence of errors in the common identifiers (the linking variables). The replicated pseudopopulations were independently randomly selected from the fictitious data on the UK population census. These data were created for the ESSnet DI (McLeod et al. 2011), which was a European project on data integration (Record Linkage, Statistical Matching, Microintegration Processing) run from 2009 to 2011. For each replicated pseudopopulation, the two lists were randomly generated according to the following coverage probabilities, $\tau_1 = 0.930$ and $\tau_2 = 0.924$, respectively.

Finally, on each replicated setting, the two lists were linked assuming three different scenarios to reflect different levels of informativeness in the linking variables. The Gold scenario uses linking variables with the highest identifying power, namely, *Name, Surname, Complete date of birth*. In this scenario, of course, the best results in terms of linked pairs and expected linkage errors are achieved.

The Silver scenario represents a situation where the strongest identifying variables – namely, *Name* and *Surname* – are not available, because, for instance, they are not released due to privacy issues. The linkage procedure can still be applied on variables with lower identification power than in the Gold Scenario, namely, the *Complete Date of Birth*. This causes linkage errors higher than in the previous scenario, in terms of both missing links and false links.

Table 4. Distribution of the linkage errors in the three scenarios

Scenario	Linkage results	Min	Q1	Median	Mean	Q3	Max
Gold	α	0.838	0.933	0.940	0.939	0.945	0.961
	β	0	0	0	0.001	0	0.057
Silver	α	0.807	0.842	0.853	0.851	0.861	0.884
	β	0.028	0.077	0.099	0.101	0.125	0.179
Bronze	α	0.808	0.822	0.833	0.833	0.843	0.874
	β	0.037	0.084	0.108	0.107	0.132	0.209

Finally, the Bronze scenario is the most unfavourable in terms of linkage errors; the set of variables used in the linkage procedure, namely *Surname, Day and Month of Birth*, has the lowest identifying power. In fact, in our data these variables are the ones most affected by typos and missing values. More precisely, in both lists, 16.7%, 2.6% and 4.3% of the records are affected by error in *Surname*, *Day of Birth* and *Month of Birth* respectively.

All the probabilistic record-linkage procedures were applied by means of the software RELAIS (see RELAIS 2011), according to the Fellegi and Sunter model summarised in Subsection 3.1.

Table 4 summarises the linkage results in terms of linkage errors, reporting the probability of nonmissing true matches (α) and the probability of false matches (β) as defined in Subsection 3.1. The true values of α and β can be evaluated in light of the true linkage status, which is known for each pair in each replication of the three scenarios.

4.2.2. Performance of the Alternative Estimators in the Simulation Study

From each linked set, we computed the counts x^*_{11}, x^*_{12} and x^*_{21} to apply the naïve Petersen estimator and the adjusted DF and MDF estimators described in Subsection 3.2 and 3.3, respectively. The DF and the MDF estimators are computed using the true values of the probability of nonmissing true matches (α) and the probability of false matches (β) obtained in each replication. The use of the true values of α and β allows the comparison of the estimators without the effect of linkage-error estimation.

To assess their performance, alternative estimates for each replicate in the three scenarios are reported in Figures 1–3.

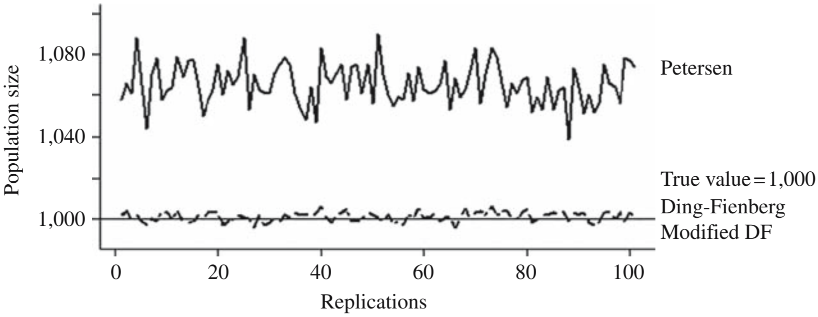

Fig. 1. Estimates in the Gold Scenario

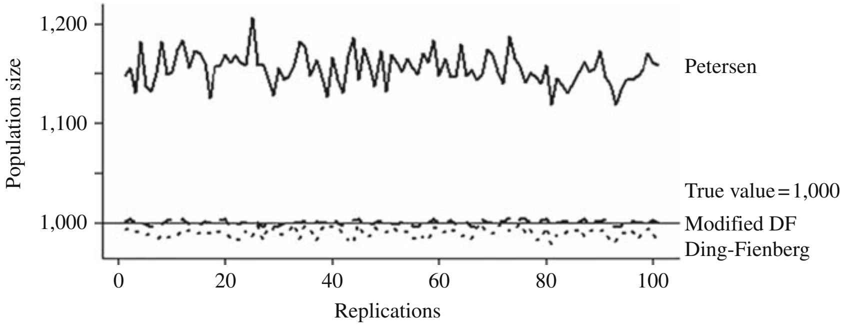

Fig. 2. Estimates in the Silver Scenario

In the Gold scenario, mimicking a situation where false linkage error is (nearly) absent, the adjusted estimators improve the naïve Petersen estimator in terms of bias as already shown with the real data application (Subsection 4.1). Again, as expected, the DF and the MDF are very close, as the extension in the MDF involves only the false linkage error β, as it results from a comparison of Equations (7) and (13) by simple algebra.

In the Silver scenario, where the false linkage error β is not negligible, the outperformance of the MDF with respect to the alternative estimators is clear. The comparison of Graphs 1–3 shows that the improvement by the MDF estimator is even more evident with higher levels of linkage error, as in the Bronze scenario.

The adjusted estimators' outperformance in terms of relative errors with respect to the naïve Petersen estimator is also shown in Table 5, where the minimum, the first quartile, the median, the mean, the third quartile and the maximum of the Percentage Relative Error over the 100 replications are reported for the three scenarios.

5. Concluding Remarks and Future Work

This work proposes a method for evaluating the unknown size of a population in the Petersen framework when the record linkage is not error free. This proposal overcomes the limitations of the Ding and Fienberg (1994) model tailored on the population census

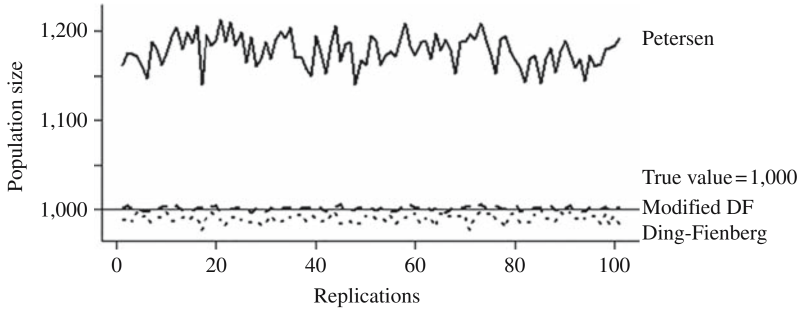

Fig. 3. Estimates in the Bronze Scenario

Table 5. Percentage Relative Error distribution in the three scenarios

		Percentage Relative Error					
Scenario	Estimator	Min	Q1	Median	Mean	Q3	Max
Gold	Petersen	3.9	5.9	6.4	6.5	7.3	9.0
	DF	− 0.4	− 0.1	0.1	0.1	0.3	0.6
	MDF	− 0.4	− 0.1	0.1	0.1	0.3	0.6
Silver	Petersen	11.8	14.5	15.6	15.5	16.4	20.6
	DF	− 2.0	− 1.3	− 0.9	− 0.9	− 0.6	0
	MDF	− 0.4	− 0.1	0.1	0.1	0.3	0.6
Bronze	Petersen	14.0	16.7	17.9	17.8	19.0	21.3
	DF	− 2.0	− 1.4	− 1.0	− 1.0	− 0.7	0
	MDF	− 0.4	− 0.1	0.1	0.1	0.3	0.6

coverage context. The application on real data showed an improvement of all the considered alternative methods in terms of bias with respect to the Petersen estimator. In this particular case, the model value of β was zero. When dealing with administrative data, this value is justified if personal identifying codes are available. In this case, the missed links are the most serious issue, since the omitting or erroneous reporting of identifying variables is not uncommon in administrative sources, in particular when they contain reference units and variables that differ from the statistical ones.

The simulation on fictitious data confirms the results of the real data application under more general frameworks, where different linkage-error levels are considered. Moreover, simulation results indicate that the MDF outperforms the other estimators when β is not negligible.

The adjusted methods depend on the correct evaluation of both kinds of linkage errors. This clearly appears in the real data application. In this application, the estimators' performances are assessed in both the following cases: linkage errors are estimated from the linkage model (Formulas 4 and 5); and the true linkage errors values are available. However, the adjusted estimators' improvement can also be observed with respect to the Petersen estimator in the first case. Further improvement in adjusting for linkage errors could be achieved by introducing individual values for the probability of correct links and missing links.

The evaluation of linkage errors is still an unresolved issue. The proposals that consider the linkage errors in analyses of linked data are often based on a training set to assess linkage quality. In any case, automatic probabilistic methods are necessary, particularly for detecting missing-link errors.

Moreover, a method for estimating the variance of the adjusted estimator is also needed. An interesting topic for future research would be the assessment of the trade-off between the gain in bias and the efficiency loss when linkage errors have to be estimated.

Finally, the effect of linkage-error adjustment should be studied for the extended models already proposed in the literature (see Section 2 for a short review) to overcome the other assumptions of the Petersen model.

Appendix

Conditions for admissibility of MDF

By straightforward algebra, estimates of the capture probabilities in (11) and (12) are positive under the following conditions for the linkage errors β and (1-α):

$$\text{a1)}\ x_{11}^*(1-2\beta) > \beta\left(x_{21}^* + x_{12}^*\right)\ \text{and}\ 2\beta - \alpha < 0,\ \text{i.e.,}\ \beta < x_{11}^*/(N_1 + N_2)\ \text{and}\ 2\beta - \alpha < 0$$

or

$$\text{a2)}\ x_{11}^*(1-2\beta) < \beta\left(x_{21}^* + x_{12}^*\right)\ \text{and}\ 2\beta - \alpha > 0,\ \text{i.e.,}\ \beta > x_{11}^*/(N_1 + N_2)\ \text{and}\ 2\beta - \alpha > 0.$$

In practical situations, the probability of linking false pairs, β, is close to zero, whereas probability of recognizing true links, α, is close to one, hence condition a1) will hold in common linkage contexts.

Furthermore, estimates of the capture probabilities in (11) and (12) are less than 1 under the following conditions for the linkage errors β and (1-α):

$$\text{b1)}\ x_{12}^* < x_{21}^*\ \text{and}\ < \frac{x_{11}^* - \alpha x_{11}^* - \alpha x_{12}^*}{x_{21}^* - x_{12}^*},\ \text{which in practice may hold only when}\ \alpha < \frac{x_{11}^*}{N_1}$$

or, on the contrary,

$$\text{b2)}\ x_{12}^* > x_{21}^*,\ \text{then}\ \beta > \frac{-x_{11}^* + \alpha x_{11}^* + \alpha x_{12}^*}{x_{12}^* - x_{21}^*},\ \text{which in practice may hold only when}\ \alpha > \frac{x_{11}^*}{N_1}$$

or

$$\text{b3)}\ x_{12}^* = x_{21}^*,\ \text{when}\ \alpha < \frac{x_{11}^*}{N_2} = \frac{x_{11}^*}{N_1},\ \text{i.e.,}\ \alpha < \hat{\tau}_1 = \hat{\tau}_2$$

6. References

Agresti, A. 1994. "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort". *Biometrics* 50: 494–500.

Bartolucci, F. and A. Forcina 2006. "A Class of Latent Marginal Models for Capture-Recapture Data With Continuous Covariates". *Journal of the American Statistical Association* 101: 786–794, Doi: http://dx.doi.org/10.1198/073500105000000243.

Belin, T.R. and D.B. Rubin 1995. "A Method for Calibrating False-Match Rates in Record Linkage". *Journal of the American Statistical Association* 90: 694–707. Doi: http://dx.doi.org/10.1080/01621459.1995.10476563.

Bock, R.D. 1975. *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Chao, A. 2001. “An Overview of Closed Capture-Recapture Models”. *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158–175. Doi: http://dx.doi.org/10.1198/108571101750524670.

Chen, Z. and L. Kuo 2001. “A Note on the Estimation of the Multinomial Logit Model with Random Effects”. *The American Statistician* 55: 89–95. Doi: http://dx.doi.org/10.1198/000313001750358545.

Cormack, R.M. 1989. “Log-Linear Models for Capture-Recapture”. *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Coull, B.A. and A. Agresti 1999. “The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies”. *Biometrics* 55: 294–301. Doi: http://dx.doi.org/10.1111/j.0006-341X.1999.00294.x.

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

Fellegi, I.P. and A.B. Sunter 1969. “A Theory for Record Linkage”. *Journal of the American Statistical Association* 64: 1183–1210. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Fienberg, S.E. 1972. “The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables”. *Biometrika* 59: 591–603. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Ghosh, S.K. and J.L. Norris 2005. “Bayesian Capture-Recapture Analysis and Model Selection Allowing for Heterogeneity and Behavioral Effects”. *NCSU Institute of Statistics, Mimeo Series* 2562: 1–27. Doi: http://dx.doi.org/10.1198/108571105X28651.

Jaro, M. 1989. “Advances in Record Linkage Methodology as Applied to Matching the 1985 Test Census of Tampa, Florida”. *Journal of American Statistical Association* 84: 414–420. Doi: http://dx.doi.org/10.1080/01621459.1989.10478785.

Large, A., J. Brown, O. Abbott, and A. Taylor 2011. “Estimating and Correcting for Over-Count in the 2011 Census”. *Survey Methodology Bulletin* 69: 35–48.

Larsen, M.D. and D.B. Rubin 2001. “Iterative Automated Record Linkage Using Mixture Models”. *Journal of the American Statistical Association* 96: 32–41. Doi: http://dx.doi.org/10.1198/016214501750332956.

Lincoln, F.C. 1930. *Calculating Waterfowl Abundance on the Basis of Banding Returns* 118: United States Department of Agriculture Circular, 1–4.

McFadden, D. 1974. “Conditional Logit Analysis of Qualitative Choice Behavior”. In *Frontiers in Econometrics*, edited by P. Zarembka, 105–142, New York: Academic Press.

McLeod, P., D. Heasman, and I. Forbes 2011. *Simulated Data for the on the Job Training*, Essnet DI. Available at: http://www.cros-portal.eu/content/job-training (accessed 20 July, 2015).

Petersen, C.G.J. 1896. “The Yearly Immigration of Young Plaice Into the Limfiord From the German Sea”. *Report of the Danish Biological Station* 6: 5–84.

RELAIS. 2011. *User’s Guide Version 2.2* Available at: https://joinup.ec.europa.eu/software/relais/asset_release/relais-221 (accessed 20 July, 2015)

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data". *Journal of the American Statistical Association* 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zwane, E. and P. van der Heijden (2005). "Population Estimation Using the Multiple System Estimator in the Presence of Continuous Covariates". *Statistical Modelling* 5: 39–52. Doi: http://dx.doi.org/10.1191/1471082X05st086oa.

Received February 2014
Revised December 2014
Accepted January 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 431–451, http://dx.doi.org/10.1515/JOS-2015-0026

Models for Combining Aggregate-Level Administrative Data in the Absence of a Traditional Census

Dilek Yildiz[1] *and Peter W.F. Smith*[1]

Administrative data sources are an important component of population data collection and they have been used in census data production in the Nordic countries since the 1960s. A large amount of information about the population is already collected in administrative data sources by governments. However, there are some challenges to using administrative data sources to estimate population counts by age, sex, and geographical area as well as population characteristics. The main limitation with the administrative data sources is that they only collect information from a subset of the population about specific events, and this may result in either undercoverage or overcoverage of the population. Another issue with the administrative data sources is that the information may not have the same quality for all population groups. This research aims to correct an inaccurate administrative data source by combining aggregate-level administrative data with more accurate marginal distributions or two-way marginal information from an auxiliary data source and produce accurate population estimates in the absence of a traditional census. The methodology developed is applied to estimate population counts by age, sex, and local authority area in England and Wales. The administrative data source used is the Patient Register which suffers from overcoverage, particularly for people between the ages of 20 and 50.

Key words: Combining data; log-linear model with offset; administrative data; England and Wales; population estimates.

1. Introduction

The population information typically collected by censuses is essential (for governments) in terms of developing policies, providing public services, and conducting research in many different areas. Censuses are used to produce population statistics (population count and characteristics) for a particular area at a given point in time. Traditional censuses, defined as the direct enumeration of the whole population by completing census forms, are used widely, and are valuable sources in terms of producing comprehensive and detailed population information for the whole country. However, despite their advantages, traditional censuses are costly, and there has been an increasing concern that the data collected by traditional censuses become outdated a short time after the census year. Several countries (such as the Nordic countries and the Netherlands) have changed their population data collection methods in recent decades, and several others have been investigating alternative methods of census data collection and producing small-area, sociodemographic statistics (such as the United Kingdom (UK), Italy and Israel).

[1] University of Southampton, Social Statistics and Demography, Social Sciences, Southampton, SO17 1BJ, UK. Emails: d.yildiz@soton.ac.uk and p.w.smith@soton.ac.uk

Most of the censuses (traditional censuses or alternatives such as register-based censuses) aim to estimate the usual resident population. Therefore it is crucial to have a detailed usual residence definition when evaluating alternative methods. Otherwise, it is possible to miss people with more than one usual place of residence or count them more than once. Accordingly, in this research, “a usual resident of the UK is defined as anyone who, on the census date: is in the UK and has stayed or intends to stay in the UK for a period of 12 months or more, or; has a permanent UK address and is outside the UK and intends to be outside the UK for less than 12 months” (ONS 2009).

Administrative data sources are an important component of population data collection and they have been used in census data production in the Nordic countries since the 1960s. A large amount of information about the population is already collected in administrative data sources by governments. However, there are some challenges to using administrative data sources to estimate population by age, sex, and geographical area as well as population characteristics. The main limitation with the administrative data sources is that they only collect information from a subset of the population about specific events, and this may result in either undercoverage or overcoverage of the population. The coverage problem occurs either when some of the usual residents are not included in the administrative data source, or when some of the people registered in the administrative data sources are not eligible to be included in the usual resident population. Another issue with the administrative data sources is that the information may not have the same quality for all population groups. One example of this are tax records, where the information about the working-age population is expected to be more accurate and up to date than that on the retired population; or health/hospital records, where the information about children and older people is more likely to be up to date. Solving the coverage problems in the administrative data sources may be problematic, especially in countries where there is no population register which collects the basic information from the entire population.

The coverage problems and the nature of the bias and inaccuracy in the administrative sources need to be clearly understood before using administrative sources to estimate populations, and action must be taken in order to obtain accurate results. For example, the dual system estimation (DSE) approach is usually used to overcome the problem of undercoverage. The DSE approach with variations is used by the UK, Israel, the United States and Australia (ONS 2012a; 2012b). In addition, Canada uses the Reverse Record Check and Census Over-coverage Study at national level to overcome the overcoverage problem (ONS 2012a). Other alternative methods include the Bayesian approach to impose constraints on the population total used in New Zealand and the calibration method used in the Netherlands (ONS 2012a; Houbiers et al. 2003).

All of the solutions regarding correcting/adjusting an inaccurate data source require combining the inaccurate source with at least one additional data source. Data sources can be matched either at individual or at aggregate level. Some preconditions must be met before two data sources are combined at individual level (Statistics Finland 2004). These conditions are listed as: legalization, public approval, unique identification numbers, comprehensive, and reliable registers. When they are not met or at least one of the data sources is not at individual level, the combination takes place at the aggregate level. Consequently, this research aims to correct an inaccurate administrative data source by combining aggregate-level administrative data with more accurate one- or two-way marginal information from an (aggregate-level) auxiliary data source and produce accurate population estimates in the absence of a traditional census.

In this research, we assume that not all of the (higher-order) information which is now provided by the traditional census will be available in the future. Hence, we only use one- and two-way marginal information from an auxiliary data source to correct the inaccurate administrative data source. In the absence of a census, the one- and two-way marginal information could be provided from different sources. Potential sources include a coverage survey or an annual survey. We consider different log-linear models with offsets and assess their accuracy for combining an inaccurate aggregate-level administrative data source with an auxiliary data source.

We present an application using England and Wales data sources, and estimate population counts by five-year age groups, sex, and region by using different log-linear models with offsets. The models used in the application estimate the 2011 England and Wales population by combining the inaccurate Patient Register with accurate one- and two-way marginal information from the 2011 Census estimates. Subsequently, the resulting population estimates are compared to ‘gold-standard’ values, and percentage difference maps for regions are produced to present the accuracy of different models. It is also possible to use these models to estimate populations in the future by combining the register data with more recent and accurate marginal information from another auxiliary source.

Section 2 describes the methodology for combining two aggregate-level data sources by using log-linear models with offsets. Section 3 presents an application of this methodology in four subsections. In the first subsection, we introduce the data sets. The second subsection deals with the model specification. The models fitted are compared according to the percentage differences between the estimates obtained from models and gold-standard values in the third subsection; then the application section ends with a discussion. Finally, Section 4 provides a brief summary and some conclusions.

2. Method

This section presents an overview of the log-linear models with offsets which are used to combine two aggregate-level data sets in the next section.

We are interested in estimating the number of people who belong to a particular age group, sex and region. We use different unsaturated hierarchical log-linear models with offsets to combine an inaccurate administrative data source, which holds accurate higher-order association structures about the population (which is not available or accurate in the auxiliary source), and an auxiliary data source, which holds up-to-date marginal distributions and two-way marginal associations, but does not provide accurate higher-order association structures for the population (a possible reason for this may be sampling error). The two data sources are combined by using one source as the basis and by imposing the structure from the other data source using the so-called offsets. The aim is to estimate accurate and up-to-date population counts by age group, sex, and region.

Willekens (1999) demonstrated the use of a simple version of the spatial interaction model which allows the same associations between origin and destination (the associations between age group, sex, and region in our models) to be produced in the estimates as in the auxiliary data. This spatial interaction model can also be expressed as a log-linear model with an offset (Willekens 1999). Recently, log-linear models with offsets have been used to combine information from different data sources to estimate migration (Raymer and Rogers 2007; Raymer et al. 2007, 2009, 2011; and Smith et al. 2010). Raymer et al. (2007) proposed log-linear models with offsets to combine the UK National Health Service (NHS) migration data with the census migration flow data which permits the inclusion of a variable of interest which is only available in the NHS migration data. Log-linear models with offsets have also been used to combine the 1991–2007 NHS registration data with the 1991 and the 2001 censuses to model interregional ethnic migration in England by Raymer et al. (2009). Their work allows the association structure employed in the models to change over time from 1991 to 2007, while the association structure in Raymer et al. (2007) was constant over time. Smith et al. (2010) took a step forward and used log-linear models with offsets to combine three sources of data (the Patient Register Data System, the 2001 Census, and the Labour Force Survey) to estimate the migration patterns of economic activity groups over time in England.

Similarly, by employing log-linear models with offsets we ensure that the selected association structures between age group, sex, and region in the auxiliary data are transferred to the estimates so that we can correct the bias in the Patient Register. Although the log-linear models with offsets have been used recently to combine information from two or more data sources, our research differs from the previous work in two aspects. First, the main interest of this research is correcting the inaccurate or out-of-date administrative data by using information from the up-to-date auxiliary data rather than adding variables from the auxiliary data to the administrative data. Second, Raymer et al. (2007, 2009) and Smith et al. (2010) assumed that a decennial census will be available in the future, whereas we consider adjusting an inaccurate administrative data source in the absence of a census. Lastly, for the application, we are able to assess different models by comparing the fitted values obtained from the models with ‘gold-standard’ values.

Usually log-linear models are fitted by using maximum-likelihood estimation. A unique set of fitted values which are the maximum-likelihood solutions for the log-linear models both satisfy the models and match the sufficient statistics (Agresti 2013). For three-way I x J x K contingency tables with variables X, Y, and Z, the minimal sufficient statistics for the XY,Z model $\left(\log\left(\mu_{ijk}\right) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}\right)$ are $\{n_{ij+}\}$, $\{n_{++k}\}$ and for the XY,YZ model $\left(\log\left(\mu_{ijk}\right) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}\right)$ they are $\{n_{ij+}\}$, $\{n_{+jk}\}$, where $n_{ij+} = \sum_k n_{ijk}$, $n_{++k} = \sum_{ij} n_{ijk}$, and $n_{+jk} = \sum_i n_{ijk}$. The fitted values for these models can be calculated directly. Unfortunately, the solutions to likelihood equations are not always direct and easy to obtain, especially for models containing complicated association structures and offsets. However, Bishop et al. (1975, 2007) mention that the maximum-likelihood estimates for any hierarchical model can be produced by iterative fitting of the sufficient configurations.

The Newton-Raphson method or the iterative proportional fitting (IPF) algorithm can be used to solve the likelihood equations when a log-linear model does not have direct estimates (Agresti 2013). In this research we employ the IPF algorithm to produce maximum-likelihood estimates like Raymer et al. (2009) and Smith et al. (2010) because it is simpler and easier to implement, and is also more transparent than the Newton-Raphson method (Agresti 2013). The IPF algorithm originally proposed by Deming and Stephan

(1940) is also called raking, raking ratio estimation and multiplicative weighting (Bethlehem et al. 2011).

Although we use log-linear models with offsets to combine information from two aggregate-level data sources in this research, it is also possible to use other approaches. One similar approach is reweighting to adjust the initial sample weights of a data set to match (the margins of) one or several tables of auxiliary variables. When the complete population distribution of the auxiliary variables is available, this approach is usually called poststratification (Bethlehem et al. 2011). If only partial population information is available about the lower-dimensional margins of tables of auxiliary variables, it is possible to use linear or multiplicative weighting. Linear weighting uses a linear regression model to obtain correction weights by summing a number of weight coefficients. If the computation of the correction weights is by multiplying a number of weight factors, it is called multiplicative weighting and is equivalent to the IPF algorithm. Although obtaining the maximum-likelihood estimates for less complicated models such as the AS model in the application is direct, more complicated models require iterative procedures. To be consistent within the estimation procedure we employ the same estimation procedure for all models.

One drawback of the IPF algorithm is that it does not produce the parameter estimates. However, it is not a problem in this research since we are only interested in estimating the population counts. Both Bishop et al. (1975, 2007) and Agresti (2013) provided examples of the IPF algorithm to estimate cell counts in three-way tables. In addition, Willekens (1983 and 1999) provided examples of using the IPF algorithm to fit log-linear models with offsets to estimate migration flows.

Let C_{asr} denote the unknown counts from age group a, sex s, and region r from a census and let Γ_{asr} be the corresponding observed counts from an inaccurate or out-of-date administrative data source.

Assume that $C_{asr} \sim Poisson(\mu_{asr})$ and consider the saturated model for μ_{asr}:

$$\log(\mu_{asr}) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \lambda_{ar}^{AR} + \lambda_{sr}^{SR} + \lambda_{asr}^{ASR}. \tag{1}$$

In order to fit this model the complete up-to-date three-way information is needed, such as a census. However, we investigate the models to estimate the population counts in the absence of a census, and assume that the accurate three-way interaction will not be available. Therefore, fitting a saturated model is beyond the aim of the research. In this research we assume that, instead of all association structures, only one or two of the age group-sex (AS), region (R), sex-region (SR) and age group-region (AR) margins which can be obtained from other alternative sources and the population total will be available in the future.

A simple log-linear model with an offset combining an inaccurate administrative source only with the total population count information from an auxiliary source is:

$$\log(\mu_{asr}) = \lambda_0 + \log(\Gamma_{asr}). \tag{2}$$

Equation (2) can also be written as:

$$\mu_{asr} = e^{\lambda_0}\Gamma_{asr}. \tag{3}$$

The final term in Equation (2) is known and referred to as an offset which imposes the three-way association structure from the inaccurate administrative data; whereas e^{λ_0} denotes the correction factor and needs to be estimated.

Combining inaccurate or out-of-date administrative data with a valuable higher-order association structure with the marginal information from an up-to-date auxiliary source allows us to update the administrative data in order to provide more accurate population estimates. In a sense, we combine the strengths of two data sources. For this purpose, we try to estimate population counts by using as little information as possible from the auxiliary source. We envisage that in the future such auxiliary information will be available through different data sources such as annual surveys, coverage surveys or rolling surveys.

In the next section we present an application where a set of log-linear models with offsets is assessed to combine information from two data sources to estimate England and Wales population counts by age group, sex, and region.

3. An Application for Estimating the England and Wales Population

This section presents an application of the methodology for estimating the England and Wales population. The section consists of four subsections. We start by describing the data sources, and continue by presenting the model specification and the model comparison. The section ends with a discussion.

3.1. Data Sources

We use data from the Census quality assurance pack, which was publicly available on the ONS (2012d) website and which provides three-way (five-year age groups, sex, and 348 local (government) authorities) aggregate-level data tables for the England and Wales population. As mentioned above, the Patient Register 2011 (henceforth referred to as the Patient Register) and the 2011 Census estimates of usual residents (referred to as the census estimates) tables are employed in this research.

Although the 2011 Census counts are provided by the ONS, the census estimates are used as gold-standard values in this research. The reason for this preference is that, while the census counts dataset only includes the number of usual residents for whom individual details were provided in the 2011 Census process, the census estimates are produced by the ONS by adjusting the census counts for undercount, overcount and people counted in the wrong places (ONS 2012d).

The second data source is the Patient Register, which is a comprehensive data source and has the highest capacity to capture the whole population in England and Wales. It includes every person in England and Wales who is registered with a NHS General Practitioner (GP) doctor. However, estimating the population of England and Wales is not its primary purpose; moreover, according to *Beyond 2011: Administrative Data Sources Report* (ONS 2012c), the Patient Register exceeds the census estimates by 4.3% at national level, and its sex ratio exceeds the census for people aged 27 to 68. In addition, percentage differences with the census estimates are within 3% only for 41% of the local authorities. In the same report, it was also shown that the inaccuracy in the Patient Register differs by sex, age groups, and local authorities.

The ONS (2012c) listed some of the reasons for the coverage differences between the Patient Register and the usual resident population as: patients who are registered in multiple areas; duplicate NHS numbers; lags in the recording of births, deaths and migrants on the NHS Patient Register; geographical variations in data quality; and differences in definitions. Another reason for undercoverage in certain regions are the armed forces bases, which have their own medical system (Scott and Kilbey 1999; ONS 2012c). In addition, according to Scott and Kilbey (1999), people receiving only private medical care; prisoners sentenced to a term of two years or more; and patients who have been in long-stay psychiatric hospitals for a period of two years or more are not included in the Patient Register and therefore cause underestimation. Detailed information about the Patient Register, and the difference between the Patient Register and the usual resident population, is presented in the Beyond 2011 NHS Patient Register report (ONS 2012c).

Despite the fact that it is biased, the Patient Register has been used to estimate internal migration and in the quality-assuring process of the 2011 Census results (ONS 2012c). A discussion of the use of the Patient Register in estimating internal migration in England and Wales can be found in Scott and Kilbey (1999).

Its ability to provide detailed information about the population will possibly give the Patient Register a key role in population estimation in the future. Therefore, it is essential to understand the nature of its bias and inaccuracy, and to investigate whether it is possible to correct it so that it can be used in the production of population estimates in combination with more accurate data sources. Scott and Kilbey (1999) also state that estimating the resident population counts for local authority district or health-authority level by using information from the Patient Register requires significant adjustments and further research. Consequently, in this research we try to correct the inaccuracy in the register by combining aggregate-level Patient Register counts with more accurate marginal distributions or two-way marginal associations from an additional source.

For this purpose, it is useful to compare the population counts by age groups in the census estimates and in the Patient Register data sets to understand whether particular age groups are less likely to be included in the Patient Register. As expected, Figure 1a shows that there is a gap between the census estimates and the Patient Register for age groups between 20 and 50. People in younger and older age groups are more likely to be registered with only a local GP, possibly because they visit their GPs more frequently than the rest of the population. Hence there is little discrepancy between two data sets for these age groups.

Smallwood and De Broe (2009) examined the Patient Register data to understand the difference in the sex ratios in the mid-year estimates based on the 2001 Census and the previous estimates, and found that the sex ratio in the Patient Register significantly differs from a 'natural' population. In addition, Smallwood and Lynch (2010) analysed the Patient Register data in a longitudinal study to understand the difference in the area of the usual residence between the 2001 Census and the Patient Register; they noted that "men are more likely to be mis-recorded in [the] GP registers compared to women". A detailed investigation of the sex-ratio patterns in population estimates can be found in Smallwood and De Broe (2009). The current research continues by presenting the Patient Register and the census estimate sex ratios for age groups in 2011 in Figure 1b. As mentioned above, the sex ratio (male/female) of the Patient Register exceeds the census sex ratio, especially in working-age groups. According to the figure, the Patient Register sex ratio is lower than

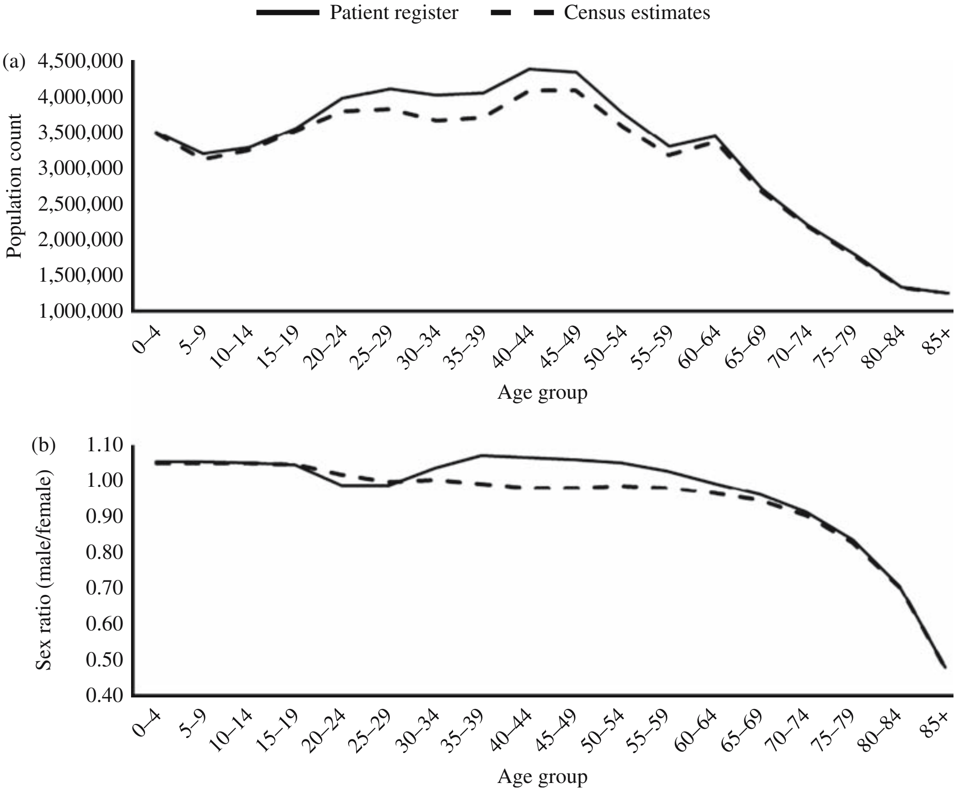

Fig. 1. (a) Population counts by age groups for England and Wales, (b) Sex ratios for the census estimates and the Patient Register (——, Patient Register; – –, Census Estimates)

the census estimates sex ratio for the 20–24 and 25–29 age groups, and it is higher than the census estimates sex ratio for age groups between 30 and 70 years old.

As we can see from Figure 1, the discrepancies between the Patient Register and the census estimates mainly occur for the population in age groups between 20 and 50 years old. In this research we aim to combine the comprehensive Patient Register, which provides biased population counts at a higher level of disaggregation, with one- or two-way marginal information from a more accurate and up-to-date data source. Here, we use census estimates data tables to provide the more accurate marginal information.

3.2. Model Specification

In this section, we present the log-linear models with offsets to estimate England and Wales population counts by 18 age groups, two sexes and 348 regions by combining the biased Patient Register with the marginal information from the accurate and up-to-date 2011 Census estimates. Note that in this application the term region refers to the 348 local authorities in England and Wales. We evaluated the total; the AS; the AS,R; the AS,SR and the AS,AR log-linear models with offsets.

The equations for these models, the equations for the IPF algorithms used in this article, which produce the same maximum-likelihood estimates, and the number of parameters in

Table 1. The IPF equations and log-linear models with offsets

Model	Explanation	IPF Equation	Log-linear models with offsets	Number of parameters
PR	The original Patient Register	$P_{asr}^{(0)} = \Gamma_{\text{asr}}$		0
Total	Adjusting grand total	$P_{asr}^{(1)} = \Gamma_{\text{asr}} \times \frac{C_{+++}}{\Gamma_{+++}}$	$\log E(P_{asr}) = \lambda_0 + \log(\Gamma_{asr})$	1
AS	Adjusting age group-sex structure	$P_{asr}^{(AS)} = \Gamma_{\text{asr}} \times \frac{C_{as+}}{\Gamma_{as+}}$	$\log E(P_{asr}) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_{as}^{AS} + \log(\Gamma_{asr})$	36
AS,R	Adding region structure to the AS model	$P_{asr}^{(AS,R)^n} = \mathrm{P}_{asr}^{(AS)} \times \frac{C_{++r}}{P_{++r}^{(AS)}}$ $P_{asr}^{(AS,R)^{n+1}} = P_{asr}^{(AS,R)^n} \times \frac{C_{as+}}{P_{as+}}$	$\log E(P_{asr}) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \log(\Gamma_{asr})$	383
AS,SR	Adding sex-region structure to the AS model	$P_{asr}^{(AS,SR)^n} = P_{asr}^{(AS)} \times \frac{C_{+sr}}{P_{+sr}^{(AS)}}$ $P_{asr}^{(AS,SR)^{n+1}} = P_{asr}^{(AS,SR)^n} \times \frac{C_{as+}}{P_{as+}}$	$\log E(P_{asr}) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \lambda_{sr}^{SR} + \log(\Gamma_{asr})$	730
AS,AR	Adding age group-region structure to the AS model	$P_{asr}^{(AS,AR)^n} = P_{asr}^{(AS)} \times \frac{C_{a+r}}{P_{a+r}^{(AS)}}$ $P_{asr}^{(AS,AR)^{n+1}} = P_{asr}^{(AS,AR)^n} \times \frac{C_{as+}}{P_{as+}}$	$\log E(P_{asr}) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \lambda_{ar}^{AR} + \log(\Gamma_{asr})$	6,282

each model are presented in Table 1. The 2011 Patient Register counts are denoted by Γ_{asr} and $P_{asr}^{(.)}$ denotes the estimated population counts for different models for age group a, sex s, and region r.

Recall that C_{asr} denote the true unobserved counts from age group a, sex s, and region r, that is, a perfect census, and that it is assumed that the census is generated from a superpopulation model where $C_{asr} \sim Poisson\left(\mu_{asr}\right)$. Note that PR denotes the original Patient Register counts. The estimates for the Total model are calculated by weighting each Γ_{asr} value by the same ratio, so that the total population estimate $\left(\sum_{asr} P_{asr}\right)$ is equal to the total census count $\left(\sum_{asr} C_{asr}\right)$. The AS model uses the age group-sex-region association structure from the Patient Register, and the age group-sex association and the total population count from the census estimates to estimate the population by age group, sex and region. The resulting estimated total population counts and the age group-sex association totals of the AS model are equal to the totals from the 2011 Census estimates. The Total and the AS models do not require iteration to fit them.

The AS,R model is constructed by adding the region structure to the AS model; the AS,SR model is constructed by adding the sex-region association structure to the AS model, and likewise the AS,AR model is constructed by adding the age group-region association structure to the AS model. In these three models (AS,R; AS,SR and AS,AR) the iteration continues until convergence is achieved. For example, for the AS,AR model the iteration continues until the marginal population totals for both the age group-sex and the age group-region are equal to the ones from the census estimates.

In this research it is assumed that the census estimates are the true values and the Patient Register is biased. The accuracies of the estimates calculated by the above models are evaluated by the percentage differences. The equation for the percentage differences for different population groups are presented in Table 2, and the comparison of the models is presented in the next subsection.

3.3. Comparison of Models

In this subsection, different log-linear models with offsets are compared according to the percentage differences between the estimates obtained from models and the census estimates. Table 3 presents the mean percentage differences between the census estimates and the estimate from different models for males and females. The mean percentage difference between the census estimates and the Patient Register without any correction for males is almost twice as high as for females. The absolute sums of percentage

Table 2. Equations of percentage differences for different population groups

Percentage differences for	Equation	Presented in
Regions	$RE_{++r}^{(.)} = \frac{P_{++r}^{(.)} - C_{++r}}{C_{++r}} \times 100$	Figure 3
Age groups for a particular sex	$RE_{as+}^{(.)} = \frac{P_{as+}^{(.)} - C_{as+}}{C_{as+}} \times 100$	Figure 2b and 2c
Age groups, sex and regions	$RE_{asr}^{(.)} = \frac{P_{asr}^{(.)} - C_{asr}}{C_{asr}} \times 100$	Figure 4 and 5

Table 3. The mean percentage differences for the Patient Register and all models for males and females

	Patient register	Total	AS	AS,R	AS,SR	AS,AR
Male	4.552	0.266	− 0.383	0.055	0.269	− 0.090
Female	2.319	− 1.875	− 0.168	0.282	0.059	0.130

differences for males and females decrease as the models increase in complexity. The smallest absolute percentage difference is achieved by the AS,R model for males and the AS,SR model for females.

The mean percentage differences for all age groups between the census estimates and different models are presented in Table A.1 in Appendix A and plotted in Figure 2a. As expected, without any corrections the highest mean percentage differences between the Patient Register and the census estimates are in the adult age groups (between 20 and 59 years). Almost all age groups in this interval have a difference higher than 3.8%. The Quality Target P1 (Maximum) mentioned in the ONS (2013) is to estimate population counts for all local authorities with a 95% confidence interval of +/− 3.8%; see Appendix B. For most of the age groups, the lowest percentage differences are for the AS,AR model. The exception to this is the older age groups. They tend to have lower percentage differences for different models.

Figure 2 shows the mean percentage differences for the age groups (a) for total population, (b) for males, and (c) for females. The mean percentage differences for the age groups for the Total model follow the same pattern as the Patient Register percentage differences, but at a lower level (not shown here). The same pattern also applies for the mean percentage differences for both males and females separately. This result is expected, since the Total model weights all the Γ_{asr} values by the same ratio $\frac{C_{+++}}{\Gamma_{+++}}$. The AS model decreases the percentage differences for almost all age groups except the youngest age group, which was already very accurate in the Patient Register. The mean percentage differences for the AS,R and the AS,SR models are very close to each other both for the total population (see Table A.1) and for males and females. Therefore, the differences for the AS,SR model are not plotted. The AS,AR model provides an almost perfect fit for the total population with the highest percentage difference of 0.06 for the 25–29 year old age group. For males, the AS,AR model overestimates the 25–29 age group and underestimates the 35–39, 40–44 and 45–49 age groups slightly. Unsurprisingly, constraining the population total to match that of the census estimates results in the underestimation of the 25–29 age group and overestimation of the 35–39, 40–44 and 45–49 age groups for females.

Considering that there are 348 regions, maps are provided for a better understanding of the effects of the models for regions. The maps present the percentage differences of the models from the census estimates for the local authorities in England and Wales. An enlarged Greater London map is also presented within the same figure since the urban areas, especially London, are subject to more internal and international migration which increases the risk of overcoverage. The maps for the total population and males in the 35–39 age group are presented in this article. The maps are divided according to the local authority quality standards specified in the ONS (2013) options paper to produce maps

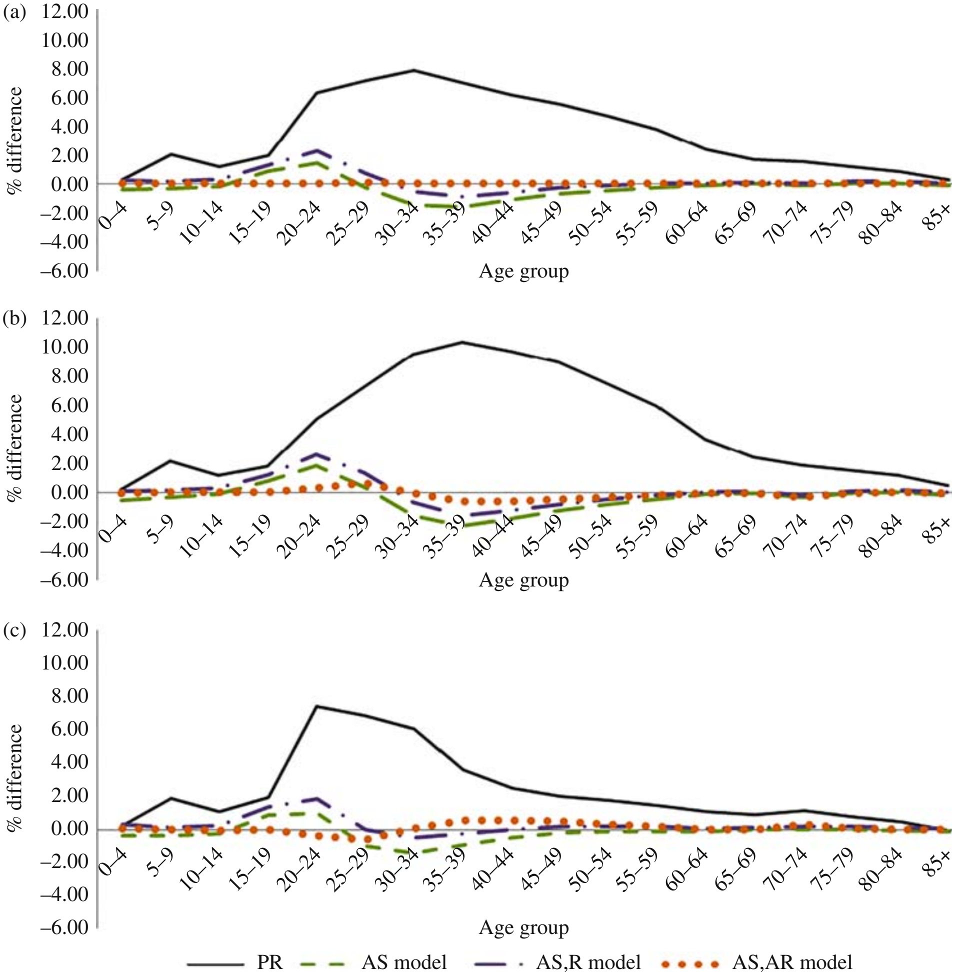

Fig. 2. Mean percentage differences according to age groups for (a) total population, (b) males and (c) females

comparable with the recent ONS publications. The ONS local authority quality standards for population estimates are presented in Table B.1 in Appendix B.

Figure 3a shows the percentage differences of the Patient Register from the census estimates. This figure shows that the Patient Register exceeds the census estimates for the total population: 57% of regions are within 3.8% of the census estimates without any correction. Figure 3b shows the percentage differences between the AS model and the census estimates for the total population: 91% of regions have population estimates within 3.8% of the census estimates after adjusting the age group-sex association structure. The remaining models (AS,R and AS,AR) are adjusting the region association in addition to the age group-sex association. Since all of the marginal region counts estimated by these models are equal to the ones in the census estimates and therefore have zero percentage differences, the maps are not presented here.

Figure 4a presents the percentage differences between the Patient Register and the census estimates for 35- to 39-year-old males. It can be clearly seen by comparing

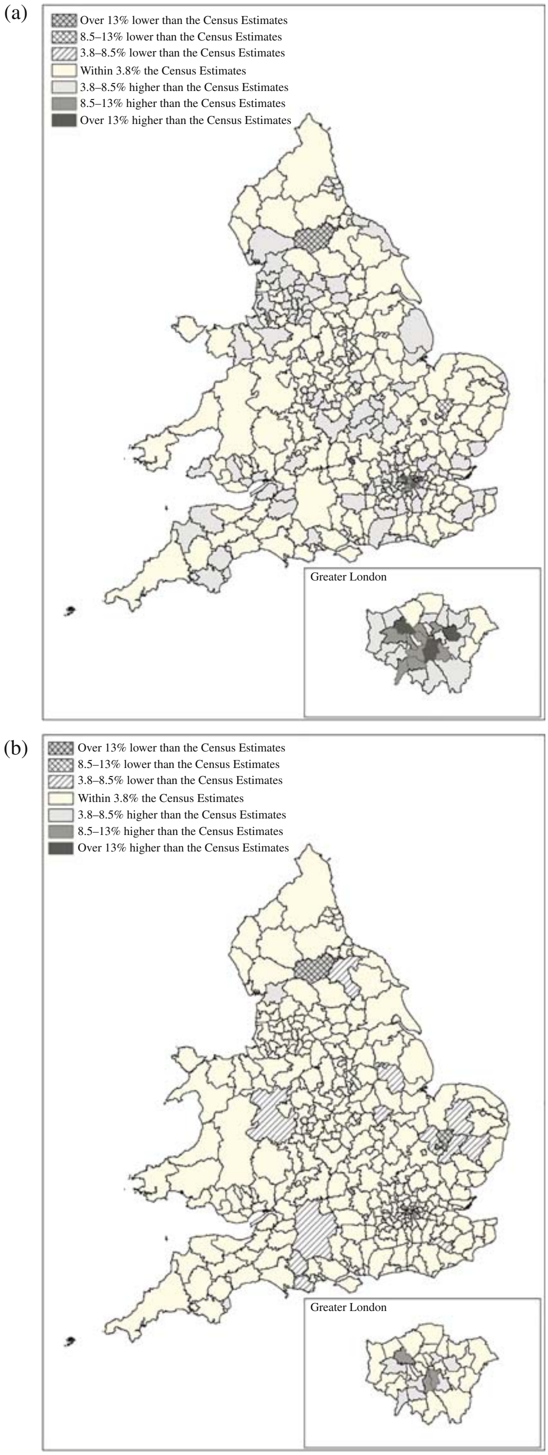
(a)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London
(b)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London

Fig. 3. Percentage differences between the census estimates and (a) the Patient Register and (b) the AS model for total population

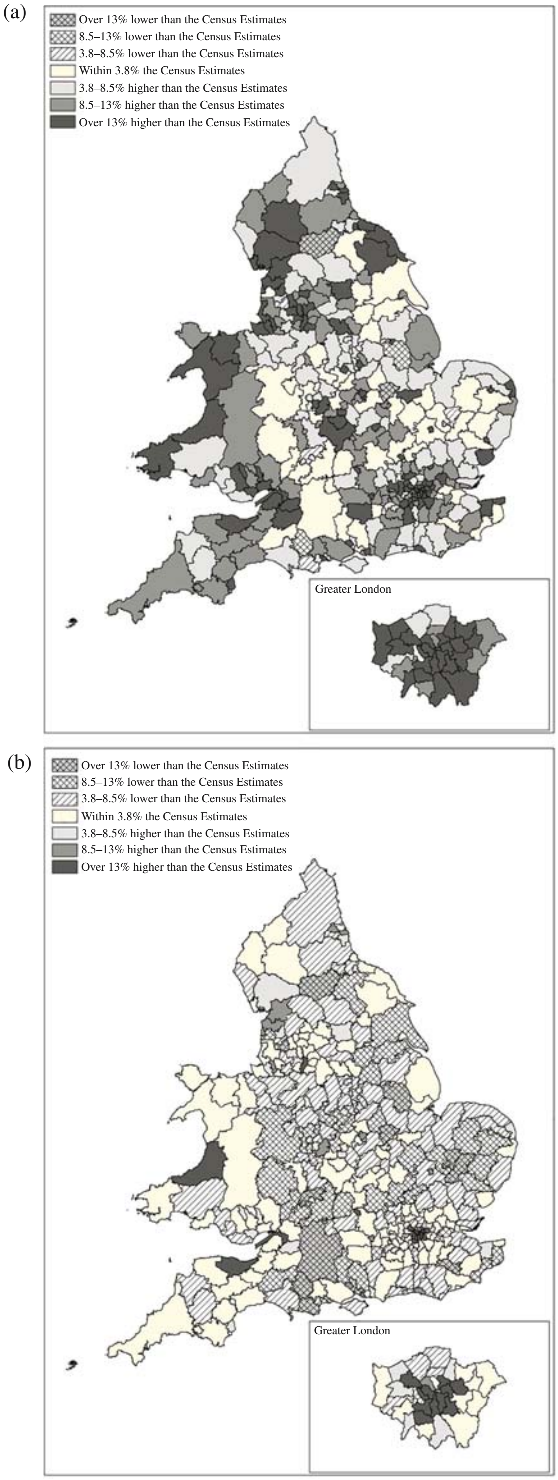
(a)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London
(b)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London

Fig. 4. Percentage differences between the census estimates and (a) the Patient Register and (b) the AS model for 35–39 males

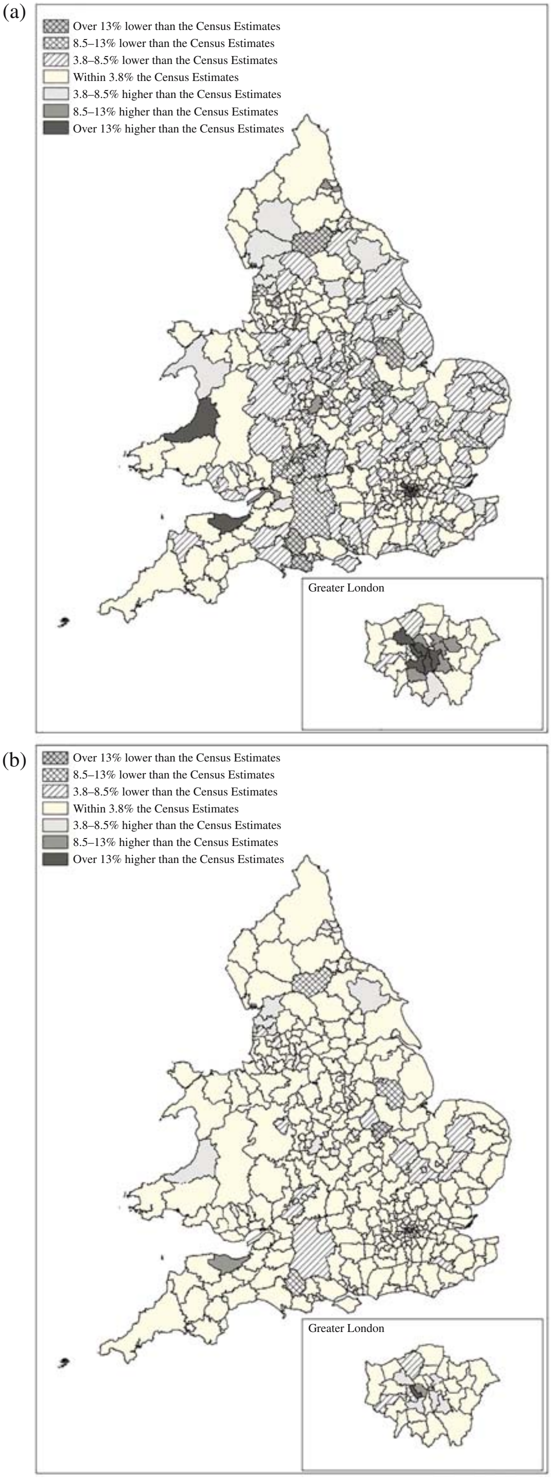
(a)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London
(b)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London

Fig. 5. Percentage differences between the census estimates and (a) the AS,R model and (b) the AS,AR model for 35–39 males

Figure 3a and Figure 4a that fewer regions are within 3.8% of the census estimates for 35–39 males than for the total population. Actually, only 16% of regions have population estimates within 3.8% of the census estimates without any correction for this age group. Comparing Figure 4a with Figure 4b is useful to see the effects of the AS model. After adjusting the age group-sex association, 39% of regions are within 3.8% of the census estimates. Taking the region association into account (AS,R model, Figure 5a) in addition to the age group-sex association (AS model, Figure 4b) improves the estimates: 52% of regions are within 3.8% of the census estimates after adjusting age group-sex and region association structures. Including the age group-region association (AS,AR model, Figure 5b) in addition to the age group-sex association dramatically improves the population estimates for the 35–39 year old males: 87% of regions are within 3.8% of the census estimates for this model.

Table 4 presents the percentage of local authorities within 3.8% of the census estimates for the Patient Register and the different models for selected age groups and sex.

To sum up, the Total model is not very effective since it weights all age, sex and region counts by the same ratio. This is not enough to solve the problems in the Patient Register. The AS model aims to correct the age group-sex structure across England and Wales and it improves the population counts to a certain extent. However, according to the percentage differences computed for the AS model, the bias in the Patient Register does not originate only from the age group-sex structure. Once the population total and the age group-sex structure are correct, adjusting the region margin also aims to correct for the overestimation and the underestimation caused by people who are not registered in their usual place of residence. The AS,R model improves the population estimates for almost all age groups. The population estimates for 35- to 39-year-old females and 40- to 44-year-old females calculated by the AS,R model are within 3.8% of the census estimates for 90% and 91% of local authorities, respectively. As expected, including the sex-region association does not dramatically improve the AS,R model since the sex distribution across the geography does not tend to change much, except for the local authorities with large armed forces bases. The smallest percentage differences for most of the age groups for males are obtained by the AS,AR model. However, it does not provide the best estimates for the older age groups (see Figure 2b).

The success of the AS,AR model indicates that not only the age group-sex association but also the age group-region association in the Patient Register requires adjustment.

Table 4. Percentage of local authorities within 3.8% of the census estimates

	PR	Total	AS	AS,R	AS,SR	AS,AR
Total population	57	88	91	100	100	100
20–24 Males	23	32	39	28	28	76
35–39 Males	16	34	39	52	58	87
40–44 Males	12	42	43	59	67	85
70–74 Males	67	45	82	79	74	97
20–24 Females	24	42	52	49	52	76
35–39 Females	57	66	66	90	87	86
40–44 Females	72	64	86	91	95	86
70–74 Females	78	34	83	82	89	98

In a typical local authority it is expected that the age distribution follows the same pattern as the England and Wales total. However, for some local authorities the age distribution may slightly or sometimes even substantially differ from the total distribution. The local authorities with universities and industrial areas with more job opportunities may attract the younger generation, whereas retired people may be keener to live in certain local authorities. To understand the difference between local authorities in detail, the pull and push factors in migration should be looked at, something which is beyond the scope of the current research.

Consequently, there is no single model that provides the best population estimates for all age and sex groups. According to our research, the AS,AR model seems to be the one which produces the most reasonable estimates. Nevertheless, the drawback of this model is that it requires both the age group-sex and the age group-region association structures to correct the bias in the Patient Register. If these association structures can be drawn from a future source, the AS,AR model might be expected to result in population estimates within 3.8% of the census estimates for more than 75% of local authorities for the five-year age groups by sex.

3.4. Discussion

The most comprehensive administrative source in England and Wales is the NHS Patient Register which covers everyone registered with a GP. As mentioned above, it is known that the direct estimates from the Patient Register exceed the census estimates. The aim of this application is to understand the nature of the Patient Register's bias and inaccuracy, and to investigate if it is possible to correct it so that it can be used as a proxy for the traditional census after being combined with more accurate data sources. We tried to correct the bias in the Patient Register by using the marginal distribution and two-way marginal information from the 2011 Census estimates. According to our research, the most effective model to decrease the discrepancy between the Patient Register and the census estimates is the AS,AR model. It improves the Patient Register in terms of percentage differences. However, it is possible that more complicated models with more marginal information might result in better estimates than our models. However, this would conflict with the aim of this research since they require more information.

In addition to the Patient Register, we also used log-linear models with offsets to correct the bias in the School Census (for age groups 5–9 and 10–14) and the Social Security and Revenue Information (for age groups younger than 15 years and older than 65 years) by using marginal information from the census estimates. However, these models did not result in better estimates than the corrected Patient Register. Accordingly, they are not presented here.

To understand the differences between the Patient Register and the census estimates, and investigate how (aggregate-level) Patient Register data can be used to estimate the population in England and Wales, further analysis of the administrative data sources is needed. This work includes but is not limited to the following: to investigate the impact of the presence of the armed forces on administrative data sources (ONS 2012c), and to investigate the impact of migration and non-UK-born residents registering/deregistering with a GP and updating their address information.

4. Conclusion

The use of already collected data for population estimation is an alternative to the costly and quickly outdated traditional census. Administrative data sources have collected comprehensive information from the population. However, they are usually not designed to collect information from the whole population and they are subject to both under– and overcoverage. Moreover, the information they collected may be biased or outdated. In the absence of a traditional census, accurate marginal information from an additional data source such as a rolling survey, annual survey or a coverage survey can be used to correct the coverage problems in an administrative data source.

This research presents a methodology to adjust an inaccurate administrative data source by combining it with an additional data source holding accurate marginal information in the absence of a traditional census. It also presents an assessment of some log-linear models with offsets in the application section according to their success in estimating the England and Wales population by age group, sex and region.

This research provides a reproducible procedure that will allow future users to estimate population counts by combining different aggregate-level sources, and it is also possible to modify the procedure to use sources with wider or narrower age bands rather than five-year age bands. In addition, this research can be extended in such a way that known issues about particular administrative sources and expert knowledge can be taken into account to develop different models. Employing the best models for different age groups and sex and combining the resulting estimates to produce accurate population counts is also possible.

Another possible approach, currently being investigated, is to use Bayesian methods to combine information from an auxiliary source with the administrative data to obtain up-to-date estimates. A recent example of combining administrative sources to estimate population counts by using a Bayesian approach is work carried out in New Zealand. Bryant and Graham (2013) produced population estimates by combining information from multiple data sources for six regions of New Zealand. However, one possible problem with this approach is a long computational time when estimating the population counts for a large number of regions and age groups, such as in our application.

Appendix A

Table A.1. The mean percentage differences for the Patient Register and all models for each age group

Age groups	PR	Total	AS	AS,R	AS,SR	AS,AR
0–4	0.229	−3.878	−0.434	0.211	0.213	0.009
5–9	2.036	−2.145	−0.353	0.176	0.189	0.014
10–14	1.175	−2.970	−0.160	0.273	0.257	0.011
15–19	1.923	−2.248	0.831	1.305	1.293	0.032
20–24	6.307	1.951	1.442	2.243	2.236	0.014
25–29	7.138	2.742	−0.273	0.713	0.706	0.064
30–34	7.873	3.447	−1.490	−0.542	−0.544	0.050
35–39	7.055	2.662	−1.578	−0.896	−0.895	0.010
40–44	6.200	1.842	−1.107	−0.618	−0.614	−0.015
45–49	5.553	1.222	−0.681	−0.283	−0.290	0.028
50–54	4.695	0.399	−0.459	−0.101	−0.100	0.007
55–59	3.764	−0.494	−0.278	0.007	0.007	0.027
60–64	2.407	−1.795	−0.111	0.050	0.052	0.014
65–69	1.690	−2.483	−0.010	0.111	0.109	0.016
70–74	1.542	−2.625	−0.135	0.023	0.003	0.013
75–79	1.182	−2.958	−0.009	0.156	0.148	0.027
80–84	0.829	−3.296	−0.016	0.161	0.156	0.031
85+	0.240	−3.855	−0.136	0.042	0.027	0.017

Appendix B

The ONS evaluates alternative options according to the population estimates quality standards achieved for the 2011 Census. Three quality standards are adopted (ONS 2013):

P1 (Maximum) corresponds to the peak-level accuracy achieved by the 2011 Census, which is that population estimates for all local authorities had a 95% confidence interval of ±3.8% or better.

P2 (Variable) corresponds to "the accuracy of the mid-year population estimates in the middle of the decade, 2006", which is that all LA population estimates had a 95% confidence interval of ±8.5% or better.

P3 (Average) corresponds to "the accuracy of the mid-year population estimates at the end of the decade, just before the next census is taken", which is that all LA population estimates had a 95% confidence interval of ±13% or better.

Table B.1. Local Authority quality standards for population estimates

Quality standard	97% of LA population estimates have a 95% confidence interval of . . .	All LA population estimates have a 95% confidence interval of . . .
P1	+/− 3.0% or better	+/− 3.8% or better
P2	+/− 3.0% or better in the peak year +/− 6.0% or better in year nine	+/− 3.8% or better in the peak year +/− 13.0% or better in year nine
P3	+/− 5.2% or better	+/− 8.5% or better

Source: ONS, 2013, Table A1: LA quality standards for population estimates

5. References

Agresti, A. 2013. *Categorical Data Analysis*. New Jersey: John Wiley & Sons, Inc.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975, 2007. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press, reprinted by Springer in 2007.

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

Bryant, J.R. and P.J. Graham. 2013. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis* 8: 591–622.

Deming, W.E. and F.F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *The Annals of Mathematical Statistics* 11: 427–444.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen, and V. Snijders. 2003. *Estimating Consistent Table Sets: Position Paper on Repeated Weighting*. Statistics Netherlands, Discussion paper 03005.

Office for National Statistics. 2009. *Final Population Definitions for the 2011 Census*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/2011-census-questionnaire-content/final-population-definitions-for-the-2011-census.pdf (accessed November 2014).

Office for National Statistics. 2012a. *Beyond 2011: A Review of International Approaches to Estimating and Adjusting for Under- and Over-Coverage*. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/research-reports/beyond-2011—a-review-of-international-approaches-to-estimating-and-adjusting-for-under–and-over-coverage.pdf (accessed June 2014).

Office for National Statistics. 2012b. *Beyond 2011: Exploring the Challenges of Using Administrative Data*. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011—exploring-the-challenges-of-using-administrative-data.pdf (accessed June 2014).

Office for National Statistics. 2012c. *Beyond 2011: Administrative Data Sources Report: NHS Patient Register*. Office for National Statistics. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/sources-reports/beyond-2011–administrative-data-sources-report–nhs-patient-register–s1-.pdf (accessed January 2014).

Office for National Statistics. 2012d. *2011 Census Quality Assurance Pack Data Tables*. Office for National Statistics. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/local-authority-quality-assurance/2011-census-quality-assurance-pack-data-tables.xls (accessed January 2014).

Office for National Statistics. 2013. *Beyond 2011: Options Report 2*. Office for National Statistics. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-options-report-2–o2-.pdf (accessed December 2014).

Raymer, J. and A. Rogers. 2007. "Using Age and Spatial Flow Structures in the Indirect Estimation of Migration Streams." *Demography* 44: 199–223. Doi: http://dx.doi.org/10.1353/dem.2007.0016.

Raymer, J., G. Abel, and P.W.F. Smith. 2007. "Combining Census and Registration Data to Estimate Detailed Elderly Migration Flows in England and Wales." *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 170: 891–908. Doi: http://dx.doi.org/10.1111/j.1467-985X.2007.00490.x.

Raymer, J., P.W.F. Smith, and C. Guilietti. 2009. "Combining Census and Registration Data to Analyse Ethnic Migration Patterns in England from 1991 to 2007." *Population, Space and Place* 17: 73–88. Doi: http://dx.doi.org/10.1002/psp. 565.

Raymer, J., J. de Beer, and R. van der Erf. 2011. "Putting the Pieces of the Puzzle Together: Age and Sex-specific Estimates of Migration amongst Countries in the EU/EFTA, 2002–2007." *European Journal of Population* 27: 185–215. Doi: http://dx.doi.org/10.1007/s10680-011-9230-5.

Scott, A. and T. Kilbey. 1999. "Can Patient Registers Give an Improved Measure of Internal Migration in England and Wales?" *Population Trends* 96: 44–56.

Smallwood, S. and S. De Broe. 2009. "Sex Ratio Patterns in Population Estimates." *Population Trends* 137: 41–50.

Smallwood, S. and K. Lynch. 2010. "An Analysis of Patient Register Data in the Longitudinal Study – What Does It Tell Us About the Quality of the Data?" *Population Trends* 141: 1–19.

Smith, P.W.F., J. Raymer, and C. Guilietti. 2010. "Combining Available Migration Data in England to Study Economic Activity Flows Over Time." *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 173: 733–753. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00630.x.

Statistics Finland. 2004. *Use of Registers and Administrative Data Sources for Statistical Purposes*, Handbook, Statistics Finland, 2004.

Willekens, F. 1983. "Log-Linear Modelling of Spatial Interaction." *Papers of the Regional Science Association* 52: 187–205. Doi: http://dx.doi.org/10.1007/BF01944102.

Willekens, F. 1999. "Modelling Approaches to the Indirect Estimation of Migration Flows: From Entropy to EM." *Mathematical Population Studies: An International Journal of Mathematical Demography* 7: 239–278. Doi: http://dx.doi.org/10.1080/08898489909525459.

Received January 2014
Revised January 2015
Accepted February 2015

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 453–473, http://dx.doi.org/10.1515/JOS-2015-0027

Linkage of Census and Administrative Data to Quality Assure the 2011 Census for England and Wales

Louisa Blackwell[1], *Andrew Charlesworth*[2], *and Nicola Jane Rogers*[3]

The 2011 Census for England and Wales made extensive use of administrative data to quality assure the estimates. This included record linkage between census and administrative data. This article describes the role of record linkage in the quality-assurance process. It outlines the operational challenges that we faced and how we resolved them. Record linkage was confined to a sample within 58 carefully selected local authorities. We found characteristic patterns of under- and overcoverage in the National Health Service Patient Register, which we illustrate here with examples. Our findings may be useful in countries that, like England and Wales, do not have a comprehensive population register to draw on and that need to understand issues of coverage in their routinely collected administrative data and the use of these data to estimate populations.

Key words: Record linkage; administrative data coverage; linkage methods.

1. The Role of Administrative Data and Record Linkage in the Production of 2011 Census Estimates for England and Wales

This article describes how administrative data were used to quality assure the 2011 Census. This included record linkage between census and administrative data, which helped us to understand the discrepancies between these data that were found when aggregate-level totals were compared. Providing new insights into patterns of under- and overcoverage in the National Health Service Patient Register, this research also helped us to understand and explain why and how census estimates differ from administrative counts in particular types of local authority. We describe the methods, systems, and processes used for the linkage, and give an overview of our results and the conclusions that we drew from them. We also outline some of the operational challenges that we had to overcome. These challenges largely stemmed from the awkward reality that the research questions to be addressed by record linkage emerged during census processing and thus could not be known in advance. Our approach may be useful for other organisations and National Statistics Institutes that do not have the benefit of national population registers and that seek to understand the representativeness of routinely collected administrative data and their use in estimating the population.

[1] Office for National Statistics (ONS), Digital, Technology and Methodology, Fareham, PO15 5RR Hampshire, UK. Email: louisa.blackwell@ons.gsi.gov.uk
[2] Department for Energy and Climate Change, 3 Whitehall Place, SW1A 2AW London, UK. Email: Andrew.Charlesworth@decc.gsi.gov.uk
[3] Office for National Statistics (ONS), Population and Demography, Fareham, PO15 5RR Hampshire, UK. Email: nicola.rogers@ons.gsi.gov.uk

The 2011 Census for England and Wales aimed to count the entire population, both people and households. Asking the same questions everywhere, the census is an important source of data for comparing different parts of the country. It also underpins nonresponse weighting in a range of key national statistics produced from surveys. Ahead of the census, the Office for National Statistics (ONS) developed a comprehensive address register. This enabled ONS to add addresses and unique codes to every census questionnaire before distribution. These codes enabled ONS to keep track of paper and online returns and helped the central office to direct field staff to high-priority areas. One of ONS's strategic aims was "to maximise overall response rates and to minimise differences in response rates in specific areas and among particular population sub-groups" (Cabinet Office 2008, 23).

No census is perfect and inevitably some people and households were missed. ONS used complex statistical techniques to estimate missed people and households. This involved a coverage adjustment based on a large survey called the Census Coverage Survey (CCS), carried out independently of the census. Record linkage between the census and the coverage survey allowed ONS to estimate the population that the census had missed using Dual System Estimation methods (DSE, for more details see ONS 2012a). The estimation process incorporated a number of quality-assurance checks.

Beyond this, ONS carried out further quality assurance including a comparison of both the census counts and the estimates, which include DSE adjustments for under-coverage, against administrative sources. The aim of the quality-assurance process was to identify where further adjustments were required before the estimates could be finalised. Examples of potential issues included difficulties in data collection, data processing, or the estimation process that could lead to errors in coverage.

Extensive checking against administrative data was unprecedented for the England and Wales Census (see White et al. 2006 for usage in 2001) and involved thousands of comparisons. Where the core checks, carried out on all 348 local authorities in England and Wales, found differences that could not easily be explained, we carried out a supplementary analysis. This comprised two stages: an initial analysis at low geographic levels, and where this did not explain differences, record linkage. Ahead of the 2011 Census, ONS described when the extra quality-assurance work would be required and how it should be prioritised (ONS 2009, 2011a, 2011b). The quality-assurance approach was developed in consultation with academics, statisticians, demographers and users of census data (For further information, see ONS 2009 and 2012b). Record linkage between the administrative sources and the census was only used to investigate and understand discrepancies that could not be resolved at the aggregate level. This was the first time that ONS had used administrative microdata in this way for census quality assurance (ONS 2013a).

The approach was a 'top-down' strategy, driven by an overriding need for efficiency and timely results. The quality-assurance process for the 2011 Census was bounded by operational delivery constraints on the one hand and a desire to publish results in a timely fashion on the other. The 'window' for quality assuring the census estimates, initially local authority by local authority and then at the regional and national levels, demanded strict prioritisation. Record linkage for unresolved data anomalies was a possibility, but only for a limited number of areas. Thus record linkage was only carried out for areas where checks on the aggregate-level data highlighted the need for more detailed investigation. In this respect the approach has some parallels with ‘macroediting’ that is used to find and correct errors in survey data by considering first the impact on data aggregates (see, for example Granquist 1991). The need for flexibility and analytic agility shaped the development of the data linkage system and the processes that we used during planning and the live operation.

The primary focus for quality assurance was the main population base for outputs, the usually resident population as of census day (27 March 2011). For 2011 Census purposes, a usual resident is defined as anyone who, on census day, was in England and Wales and had stayed or intended to stay for a period of twelve months or more, or had a permanent address in England and Wales and was outside of England and Wales on census day and intended to be outside for less than twelve months. This article sets out work done by the Census Quality Assurance Data Matching team, which began in 2011 and finished in 2013.

Section 2 of this article describes the administrative sources that were available for record linkage, together with the census information that we used, in addition to census responses. In Section 3 we then discuss the operational challenges that we faced, which were dominated by the need to complete the quality-assurance process quickly in order to publish timely results. Ahead of record linkage, the administrative data were used at aggregate level to address data anomalies that the core quality-assurance processes identified. Section 4 describes what we learnt from the aggregate-level comparisons. Section 5 describes our linkage methods and we present our results for 58 local authorities in Section 6. Our conclusions, in Section 7, aim to assist other National Statistics Institutes planning to make increased use of administrative data for population estimation in a census context. We also describe how ONS is taking forward administrative record linkage to support the 2021 Census.

2. The Data Available for Linkage

The 2007 Statistics and Registration Service Act provided a legal gateway for ONS to access record-level data (microdata) from other government departments for the purpose of population estimation. Through these and other provisions, ONS gained access to the NHS General Practitioner (GP) Patient Register, the School Censuses of England and Wales, the Higher Education Statistics Agency (HESA) Student Records, the Live Births Register, the Deaths Register, Electoral Registers, and Valuation Office Agency data.

Record linkage focused primarily on the NHS Patient Register. The Patient Register includes the general identity details of patients registered with GPs. It is used within the NHS for calculating payments to GPs and for the selection of NHS patients for participation in health-screening programmes. It is one of the largest population databases in operation in England and Wales. The Patient Register was the highest-quality record-level source with the widest population coverage that was available to us at that time. We also anticipated that queries about 2011 Census estimates from key users would be based on local Patient Register counts. In addition to using the Patient Register to quality assure the census counts and estimates, we needed to understand the quality of the Patient Register and its patterns of coverage, relative to the census, to respond to stakeholder queries following the publication of census results.

Quality checks ahead of record linkage confirmed that live births and deaths were reflected accurately in the Patient Register, so these were not included in this linkage exercise.

The census data used in record linkage included: census responses (both households and individuals), including 'dummy form' information which is supplied by enumerators for nonresponding households; the census address register; the census address register history file (ARHF), which contained addresses that were assessed as nonresidential or derelict and therefore not sent a census questionnaire; census 'associated address' records, including responses to the census question 'One year ago, what was your usual address?'; second residence addresses (including students' term-time addresses) and visitors' usual residence; field operation information drawn from the Census Management Information System (CMIS); census questionnaire images. Census questionnaires have been securely destroyed, but ONS is obliged to retain census questionnaire images, which will be made publicly available in 2111.

3. Building a Linkage Methodology and Architecture for Census Quality Assurance and the Imperative for a Flexible Approach

A number of issues and uncertainties demanded a flexible approach to record linkage. Some of the challenges we faced, and their resolution, were:

3.1. Security Risks

A number of physical, technical, statistical, and legal safeguards ensured that the microdata used for linkage were handled securely. Physical safeguards included restricting their use to the census physical safe setting, where security doors ensured that only authorised staff could enter. Technical safeguards included holding and processing microdata within the census IT environment, a closed and monitored system that did not allow users to copy, print or download the data being processed. The linkage design provided statistical protection as most of the linkage was within postcodes used for the CCS, and these are not publicly known. In addition, identifying information such as name, date of birth, postcode, and address were only used to link record pairs and were not stored in analytical datasets. Legal safeguards included the requirement for all staff, including the clerical matchers, to sign the Census Confidentiality Undertaking and Declaration and receive Defence Vetting Agency Security Clearance. The penalty for a breach of data confidentiality could be a prison sentence, and all staff in the matching team signed confirmations that they understood this.

3.2. Uncertain Analytic Requirements

It was impossible to predict all of the issues that record linkage would need to address. The geography or population subgroup under consideration would determine which administrative data should be used. The data architecture therefore had to allow linkage between all or just some sources, with capacity to add new sources if they became available. 'Data architecture' refers to the collection of interlinked tables used to store the results of all address and person linkage. These were held separately for each local authority to maintain file sizes that were efficient to process. Using local authorities as the basic unit for analysis also reflected the quality-assurance process, which considered and approved the estimates for each local authority in turn.

3.3. Late Availability and Uneven Quality of Data

Only the Patient Register, Valuation Office Agency and census address register were available from the start of the quality-assurance process. Census person data became available as local authorities were processed (mirroring the order in which quality-assurance issues were raised), while CCS and other census information were only available late in the process. HESA, English and Welsh School Census and Births data became available after the quality-assurance process had begun. Electoral Register data were available for most local authorities, but were inconsistently formatted and required substantial cleaning and standardisation. A key requirement for the data-linkage architecture was the ability to incorporate new data if and when they became available.

The linkage algorithms that we used and the sequence of linking different sources had to remain flexible during the operation. For example, the School Census data for Wales were only available at a higher geographical level than for England. Our data tables and record linkage programmes were adapted to reflect this difference. Likewise, the Electoral Register cleaning and preparation revealed missing data for some local authorities. Where this occurred, we requested that records be resupplied and the subsequent delays impacted on the sequencing of local authorities through the linkage process.

3.4. The Requirement for Timely Results

Census quality assurance involved the approval of 348 local authority estimates at a series of Quality Assurance Panels (for more detail see ONS 2009 and 2011a). Where data issues could not be resolved using data at aggregate level, record linkage was used. Our systems and methods were designed to respond quickly to these requests, involving automation where possible.

The Census for England and Wales took place on March 27, 2011. ONS was committed to publishing the results in July 2012. The final agreement to publish the estimates was made by an Executive Quality Assurance Panel, the National Statistician and the Director General, executive ONS management and executive management representation from the Welsh Government. To achieve this, the estimates needed to be quality assured by April 2012. Census estimates were available for assessment by Quality Assurance Panels of ONS and external experts from September 2011. This provided just over six months to approve the estimates for all 348 local authorities, for the regions and at the national level in England and Wales. Within this brief window, record linkage, which is very labour intensive when it is supported by a clerical review of links being made, had to be done in a selective and efficient way.

To reduce the turnaround time required for data linkage results, we linked Patient Register addresses to the census address register in 37 local authorities ahead of the census. These local authorities were mostly areas of high population turnover, taking into account migration patterns since 2001. As census processing got underway, they were prioritised by the expected delivery date for their processed person-level data, in anticipation of the order they would be considered by the Quality Assurance Panels. Record linkage for each of these local authorities was suspended if they were approved by the Quality Assurance Panel, and new areas not in the original list of 37 were added as new issues arose. These included some local authorities whose estimates fell outside the tolerance bounds set for the core checks (described in ONS 2012b), and where further analysis using aggregate-level data could not resolve the anomalies. By the end of the operation, data for 58 local authorities were linked. Identifying the more challenging local authorities and completing address linkage ahead of the live operation allowed preliminary work to proceed in an intelligent way, and maximised the number of local authorities overall that could be linked.

We included a number of local authorities with stable populations, which pose little enumeration challenge because they have low levels of international and internal migration. These provided a context for the results for more challenging areas. They also validated the linkage methods that we used.

3.5. Keeping the Scale of the Linkage Task at a Manageable Level

Some quality-assurance issues were concerned with small geographic areas or population subgroups, such as students in communal establishments or babies under the age of one. Where issues were generalised across the population, linkage typically focussed on the postcodes used for the CCS (for more details, see Abbott 2009). The CCS is a sample of approximately one per cent of the country carried out after the main census and is used to create the census estimate. The CCS uses a selection of postcodes within Output Areas (OAs), which are re-enumerated independently from the census field operation. The CCS selects a sample of OAs, stratified by local authority and a national 'hard-to-count' index. Output Areas (OAs) are the lowest geographical level at which census estimates are provided. They are built from adjacent postcodes. OAs cover 40–250 households and 100–600 people and postcodes have an average of 15 households. The 'hard-to-count' index is a proxy measure for census nonresponse (for further details, see ONS 2012a). The CCS re-enumerates approximately half of the postcodes within the selected OAs and contains more postcodes in areas where the census response rate was expected to be lower. Administrative data linkage within these clusters of postcodes provided a strategic sample that constrained the scale of the record-linkage task and also provided CCS data as an additional data source for comparison against the administrative data. Crucially, by using this sample we were able to provide record linkage and analysis for a greater range of local authorities.

3.6. Ensuring Quality and Consistency in Record Linkage

The quality of record linkage, both automatic and clerical, was monitored and managed through two processes. The first involved a continuous feedback loop of linkage best practice for the clerical matching team. An example of this was the accumulation of knowledge and experience in ethnically-specific naming conventions and variations. The second involved an expert matcher's review of linkage decisions, using both a random sample and having two matchers complete the same linkage. Systematic discrepancies were addressed through further training and review.

3.7. Complexity of Linkage and Storing Results at Both Individual and Address Levels

Storing the results was complicated by the large number of sources used, the two levels at which linkage took place (addresses and individuals) and the reality of one-to-many links for both addresses and individuals. These complications meant that extra care was necessary to analyse the linkage results. One-to-many links for addresses arose from less precise recording of addresses, for example in the Patient Register. This typically involved subdivisions within buildings (for example 'Flat 1') being omitted from a Patient Register address. Thus a number of addresses in the census, referenced in more detail, could link to a Patient Register 'shell' address. One-to-many person-level links arose from multiple enumerations of individuals in the census (discussed more fully in ONS 2012c). In addition, the linkage process allowed unlinked addresses to be linked as a result of person-level linkage, for example where capture errors (a typical example was where data scanning read marks on the paper questionnaire as characters) produced address differences that confounded the address linkage the first time around.

4. What We Learnt from Comparing Different Sources at Low-Level Geographies

Core checks, applied to all local authorities, included checking estimates by age, sex and other key variables against a range of aggregated administrative and survey sources. Where the core checks identified data anomalies, supplementary checks were carried out. These involved exploring the data at a low geographical level, mainly Output Area (OA) or above. Some checks were at postcode level.

In most cases, supplementary analysis resolved apparent data anomalies. The anomalies tended to arise as a result of two main problems, the first of which is the time lag that is inherent in many of the administrative sources. People's circumstances change (for example they move house), and there is a delay before this is captured in their administrative data. The failure of most administrative systems to capture reliable, timely information on migration leads to inflated datasets containing invalid records. A second problem that we found was a degree of subjectivity in addressing. Administrative systems vary in the level of detail or accuracy used to record where people live. This was more problematic where people live in subdivided properties. Typically we found that the census information was more timely and accurate. An exception was the addressing for student halls of residence, where the census sometimes captured the administrative building that census forms were sent to for onward distribution, whereas HESA data captured students' dwellings with greater geographic accuracy.

Supplementary analyses included comparisons between the census and Patient Register at person and household level. For example, a discrepancy between census and Patient Register counts in Westminster found one area where there were several thousand more patient registrations than census individuals. Analysis by age found that the excess patient registrations were mostly of student age. Further investigation revealed that this area contained a medical centre attached to a London university. The address for this centre was wrongly given as the home address of many students registered with the practice.

In areas with high concentrations of students, the number of patient registrations often exceeded census counts for young adults. Further investigation revealed that Patient

Register counts implied that student halls were filled beyond their published capacities (see Figure 1), with the ratio of registrations to published capacities frequently higher than one. Further analysis of the date that these patients were registered confirmed that former residents had almost certainly moved on but not updated their NHS records, either because they had not yet reregistered with a new GP or had left England and Wales.

We also compared census counts and estimates against, among others: Patient Register counts of under 1's and those in the Register of Live Births; School Census counts of ethnic groups; lists supplied by local authorities of addresses containing 'annexes', along with Valuation Office Agency information and Patient Register counts; Patient Register and School Census counts for addresses within holiday parks; international migrants as defined by Patient Register records with 'flag 4' status, given to new registrations from abroad.

To understand and explain a substantial difference between census counts and council tax records, we found one area where the census found fewer than twenty households, yet the council tax data showed several hundred more. This was explained by a large block of flats that had been almost completely emptied for demolition.

5. Linkage Methods

Figure 2 summarises our record linkage processes. Linkage involved exact automated linkage, score-based automatic linkage (using similarity scores), clerical resolution of candidate pairs generated by the automatic systems, and a clerical search for residual records. This was a unique exercise carried out to validate the census.

5.1. Data Preparation

Each administrative dataset was standardised and cleaned, including removing duplicates, checking and aligning variable formats, checks for coding inconsistencies, and checking the number of unknown or missing values for each variable (1.1 in Figure 2). Electoral Registers were the most resource intensive to prepare. Maintained and supplied by

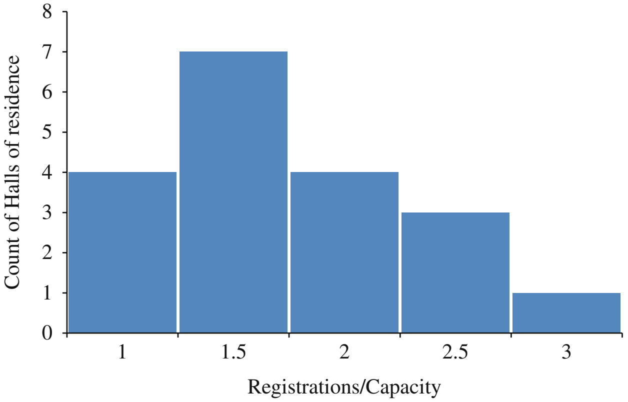

Fig. 1. Ratio of patient registrations and published capacities in student halls, in a sample of halls of residence in one university town

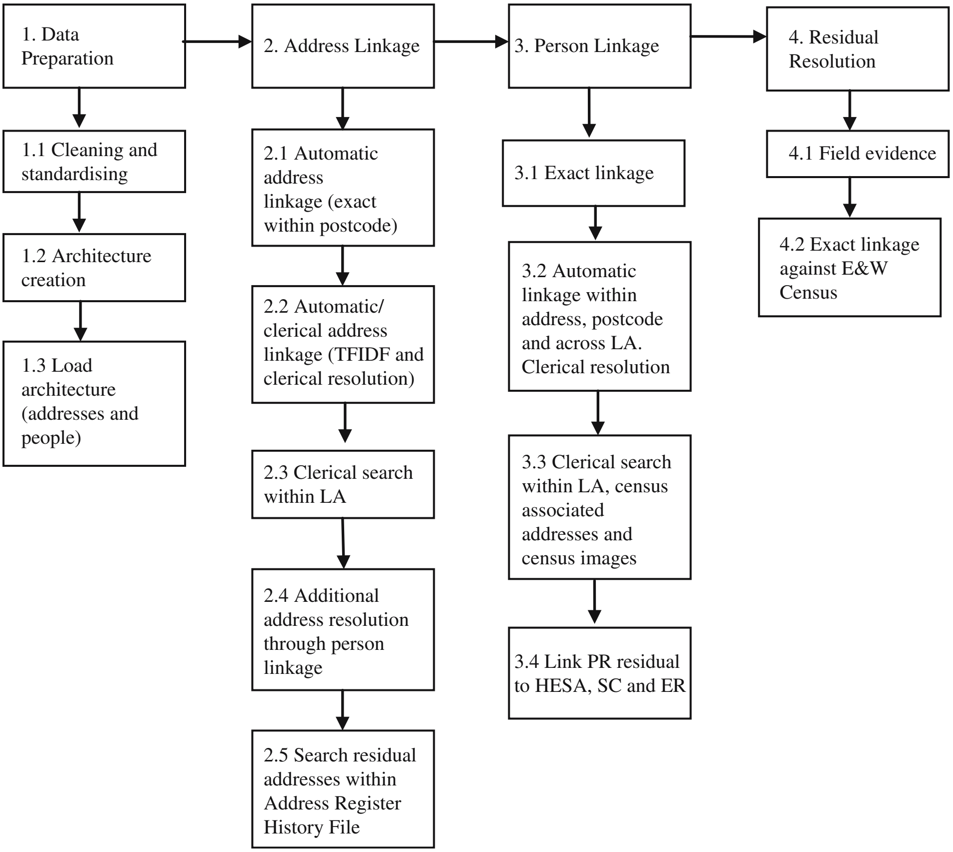
1. Data Preparation
1.1 Cleaning and standardising
1.2 Architecture creation
1.3 Load architecture (addresses and people)
2. Address Linkage
2.1 Automatic address linkage (exact within postcode)
2.2 Automatic/ clerical address linkage (TFIDF and clerical resolution)
2.3 Clerical search within LA
2.4 Additional address resolution through person linkage
2.5 Search residual addresses within Address Register History File
3. Person Linkage
3.1 Exact linkage
3.2 Automatic linkage within address, postcode and across LA. Clerical resolution
3.3 Clerical search within LA, census associated addresses and census images
3.4 Link PR residual to HESA, SC and ER
4. Residual Resolution
4.1 Field evidence
4.2 Exact linkage against E&W Census

Fig. 2. 2011 Census quality-assurance record-linkage process. Abbreviations: LA (local authority), PR (Patient Register), SC (School Census), ER (Electoral Register), E&W (England and Wales)

individual local authorities, these registers were held in a wide range of formats. Some of the standardisation could be automated, while some rare and unique differences required manual correction.

We attached geography codes to addresses using the software package ‘Matchcode’, supplied by Capscan (for more detail see http://www.capscan.com/).

We aligned the administrative sources to census definitions where possible. For example, HESA data record all students on a course at an institution within an academic year, regardless of the course duration, so individuals may have multiple instances within an institution in the same academic year. To align these data with census definitions, we used a subset of HESA records for those aged 18 and over with start dates before and end dates (or continuing) after March 27, 2011 (census day). Rules to prioritise multiple records were applied to select just one record for linkage. See ONS (2012d) for more information on the challenges of aligning definitions.

The final two stages of data preparation, ‘Architecture creation’ and ‘Load architecture’ (boxes 1.2 and 1.3 in Figure 2) refer to the creation of interlinked data tables where we stored the data for linkage and the linkage results.

5.2. Address Linkage

Addresses in the Patient Register, Electoral Register, Valuation Office Agency data, and the English School Census were linked with those in the census address register within CCS postcode clusters in selected local authorities.

Exact linkage (2.1 in Figure 2) used only flat number/property subdivision/house name, house number and road, and finally postcode. Variables with low discriminatory power such as town were excluded as they could only introduce error.

A second stage (2.2 in Figure 2) used 'Term Frequency Inverse Document Frequency' (TFIDF) linkage, which assigns a weight to each pair of words in a pair of addresses, depending on how commonly the words within the addresses appear in each of the datasets (Li et al. 2010). TFIDF linkage used all available address elements. Common terms within the address, such as 'town', calibrate and weight the less frequent ones. Linked records incorporating 'Hill Street' would have a lower weight than 'Segensworth Road', due to the rarity of 'Segensworth'. Scores for each address are weighted according to the number of words included in the address. The best-scoring candidate match for each address was referred for clerical review and confirmation.

A third stage (2.3 in Figure 2) involved a clerical matcher searching for an address match, firstly within the given postcode and then across the local authority as a whole.

Inaccuracies in recording addresses led to some addresses being falsely unlinked. Some of these addresses were subsequently linked through person linkage (2.4 in Figure 2). Where individuals living in unlinked addresses were linked, a check was made to see if these were falsely unlinked addresses due to data discrepancies.

Finally, in addition to searching for matches within census data, the Address Register History File (ARHF) was also checked (2.5 in Figure 2). The ARHF contained addresses that had not been sent a census questionnaire, for example because they were commercial addresses or known to be derelict buildings.

5.3. Person Linkage

Individuals within the Patient Register were linked to census records. Unlinked patient registrations were then searched for within the Electoral Register, School Census, and HESA data to assess the strength of their presence in administrative data. Our linkage strategy was deliberately designed to maximise linkage rates while minimising false links.

As with address linkage, the first stage of person linkage (3.1 in Figure 2) was exact linkage using forename initial, the first three characters of surname and full date of birth (dd/mm/yyyy).

A number of automatic linkage strategies followed (3.2 in Figure 2), firstly using the results from address linkage. Within linked addresses, the linkage criteria were relaxed to forename initial or a SPEDIS value of less than 100, first three characters of surname or SPEDIS value of less than 100 and two of the three date-of-birth elements matched. SPEDIS measures how close the spellings of two words are. It is a function within the SAS statistical analysis software package. The lower the score, the better the match. This relatively high threshold allowed potential matches to be referred for clerical resolution (scores of 101–200 were disallowed).

Within matched postcodes, linkage criteria were the first three characters each of forename and surname and two of three elements of date of birth. When searching more widely for CCS postcode cluster records within a local authority, forename, surname, date of birth, and sex all needed exact matching.

There then followed rules-based linkage techniques (see Li et al. 2006). Firstly, within local authorities, individuals with the same day and year of birth and sex were linked using month of birth, exact forename and surname with a qgram threshold of 0.4 or above. Qgrams measure the level of agreement between groups (in our case, pairs) of characters within the two strings being compared (the code for the qgram comparison is available from ONS upon request). The second strategy required exact surname matching and forenames with a qgram threshold of 0.4 or above.

All exact matches were recorded without further scrutiny. For individuals linked within linked addresses, those with name discrepancies, where sex was uncoded and where there was error in dates of birth were referred for clerical confirmation. All matches within postcodes and local authorities were reviewed clerically, as were duplicate matches and those identified through the rules-based linkage strategies.

Unlinked patient registrations were searched for clerically, firstly across the local authority and secondly through ‘associated address’ information (3.3 in Figure 2). This involved matching against census respondents who gave the Patient Register address as their usual address one year ago, as a second residence or as a usual residence for visitors.

To identify census matches missed because of potential data-scanning error, census form images were checked.

Where linked individuals were in addresses that were unlinked, these were referred for clerical review. In this way, addresses that either were recorded very differently between sources or contained scanning error were resolved (2.4 in Figure 2). Clerical matchers were able to carry out free text searches on name and address and any combination of day, month and year.

5.4. *Residual Resolution*

Any patient registrations that remained unlinked at the end of this thorough linkage process were searched for within the other administrative sources: the Electoral Register, School Census, and HESA data (3.4 in Figure 2).

To further resolve unlinked records, we used evidence from the census field operation (4.1 in Figure 2). Where there was no response to the census, enumerators classified addresses according to evidence they could find in the field. Thus we were able to classify unlinked records as having an address that appeared to be a second home, having an address that was occupied but the occupants were refusing to comply with the census, or as clearly vacant.

A final person linkage stage involved searching, using exact matching, across England and Wales as a whole (4.2 in Figure 2).

Table 1 provides examples of linkage rates achieved through exact, rules-based, and clerical methods. It highlights the limitations of exact linkage. Inconsistencies between names on the Patient Register and on the census form arose for a number of reasons including inconsistencies in recording names, such as abbreviations (‘William’ and ‘Bill’ for example), middle names given as forenames, inconsistent translations from

Table 1. Patient Register to 2011 Census linkage rates at each processing stage

Local authority	LA with a stable population Aylesbury Vale	Metropolitan LA Birmingham	Inner London LA Lambeth
Total number of patient registrations in the sample	2,732	21,313	10,532
% Exact linkage (3.1 in Figure 2)	54.0	50.3	34.3
% Rules–based linkage with clerical resolution (3.2 in Figure 2)	13.8	18.3	19.0
% Clerically linked (3.3 in Figure 2)	21.0	13.0	10.0
Final linkage rate	88.8	81.5	63.3

non-English (such as Chinese or Russian) characters into the English alphabet, and scanning error, among others.

6. Linkage Results for 58 Local Authorities

Record linkage proceeded on a local authority by local authority basis. Areas were selected for record linkage either because they were high migration areas where the different sources were most likely to diverge, because the Quality Assurance Panel had identified data anomalies and wanted further analysis, or because we had identified them as a useful benchmark against which to compare more challenging areas. As the number of local authorities with linked patient registrations grew, it became clear that a typology of local authorities was visible in the data.

Inevitably, not all records can be linked. Firstly, some census respondents are not registered with an NHS GP. Examples include new arrivals to England and Wales who are yet to register with a GP; those who have moved to a new area and not updated their GP registration; people using private health care rather than the NHS; those covered in the NHS outside of the GP system, such as prisoners or members of the armed forces.

Secondly, although the census aimed to capture the entire population on census night, some people were missed (the 2011 Census person response rate was 94 per cent and overcoverage was estimated at 0.6 per cent). ONS estimates the extent of undercoverage (at six per cent) using the CCS and DSE or Dual System Estimation. In terms of the administrative record linkage carried out for census quality assurance, the individuals that the census and the CCS missed could appear as unlinked patient registrations.

There is also an issue of synchronicity between the datasets. The census provides a snapshot of the population of England and Wales on census night, March 27, 2011. The Patient Register extract was taken on April 23, 2011. The gap between these reference dates was to allow people moving house to register with a GP in their new area. However, some people take longer for this so there will always be some disagreement between the sources, even in areas with relatively little population turnover (Smallwood and Lynch 2010). Moreover, if people who leave the country do not inform their GP that they are going, they remain on the register until the local health authority cleans them off the list.

Population groups that are absent or over-represented on the Patient Register produce characteristic differences in the demographic profiles for local areas. Area characteristics also shaped the patterns of coverage, as we show below for university towns. As a further example, Richmondshire and Forest Heath local authorities are home to large military bases and here the 2011 Census estimate exceeds the Patient Register count by 15 and 14 per cent, respectively. Among males aged 16–64, this rises to 37 and 15 per cent. Kensington and Chelsea have 2011 Census estimates that are six per cent higher than the Patient Register count for those aged 65 and over, reflecting a concentration of private healthcare users here (see ONS 2012e).

Powys, where the 2011 Census estimated 133,000 usual residents, had the highest Patient Register linkage rate of 93.7 per cent. This left 104 unlinked patient registrations and 231 unlinked census records *within our sampled postcodes*, where the total number of patient registrations was less than 2,000. Areas with stable populations typically had linkage rates above 85 per cent. Areas with higher levels of population turnover had Patient Register linkage rates of between 75 and 85 per cent. Linkage rates below 75 per cent typically occurred in London, and were lowest in the Inner London boroughs, where population turnover and international migration are at their highest. Kensington and Chelsea had the lowest linkage rate, with fewer than two thirds (60.5 per cent) of patient registrations linked to the census. However, comparisons of unlinked records and the coverage adjustment in each area (not shown here) provided further confidence in the census estimates. Linkage rates are summarised in Table 2.

6.1. Local Authorities With Stable Populations

Figure 3 shows the unlinked Patient Register and census records for males in a local authority with a stable population. Areas with high record-linkage rates were those with low levels of internal and international migration. Unlinked patient registrations tended to be higher for working-age people. There were more unlinked census records (dashed lines in the graphs) than unlinked patient registrations (dotted lines). This was true for most of the local authorities where linkage rates were high. Even in these areas where the two sources were most closely aligned, there were more unlinked records for men than for women (not shown here).

Patient Register records appear to be less accurate for men, who visit their GPs less frequently. This leads to longer time lags in updating NHS registrations when men move house than when women do, and the result is that Patient Register entries refer to people, men in particular, who no longer live in the area. This is more problematic in local authorities with less stable populations, such as inner-city areas, which people migrate to for work or study purposes.

6.2. Inner London

The discrepancy between the census and the Patient Register was greatest in Inner London. Figure 4 shows the linkage results for males in an Inner London local authority. For men between the ages of 25 and 44, the Patient Register had more unlinked records than records that linked to the census.

Table 2. Summary of person linkage rates

	April 2011 Patient Register			2011 Census		
	Average linkage rate* (%)	Highest linkage rate (local authority) (%)	Lowest linkage rate (local authority (%)	Average linkage rate* (%)	Highest linkage rate (local authority) (%)	Lowest linkage rate (local authority (%)
All 58 local authorities	79.7	93.7	60.5	81.2	93.6	63.9
Local authorities with stable populations	88.3	93.7	80.9	87.5	93.6	77.8
Metropolitan areas excluding Inner London	81.0	85.6	76.9	80.0	88.1	73.5
Inner London	68.3	75.2	60.5	72.3	79.2	63.9

*This is measured across all local authorities or local authorities in this category and is the total of the individual linkage rates divided by the number of local authorities in that category. For each local authority, the Patient Register or census linkage rate is the number of *linked* registrations or census records as a percentage of the *total number* of registrations or census records.

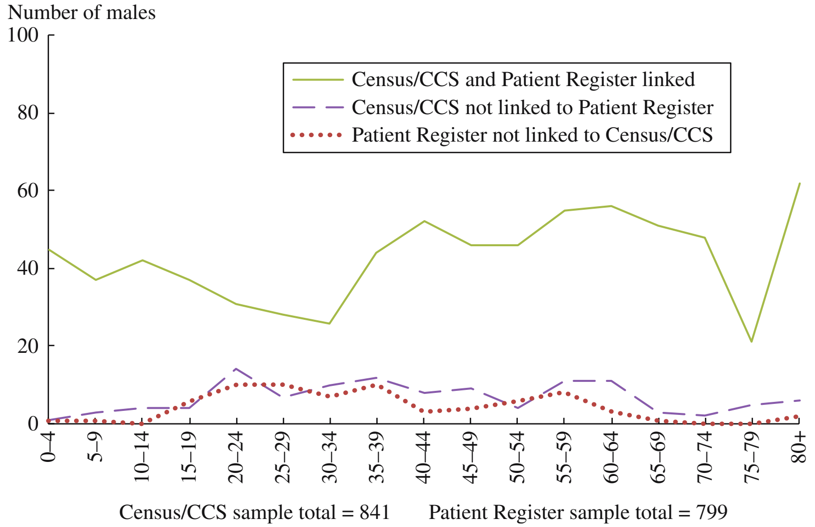

Fig. 3. Census and Patient Register linkage results for males in an area with a stable population, by age

6.3. University Towns

Another pattern that we found in some areas with high proportions of students is illustrated in Figure 5. Here, there were more unlinked census records for men aged 20–24 than were linked to the Patient Register. Thus over half of the men in this age group that the census or CCS captured were different to those on the local Patient Register. Many students leave their patient registrations at their home (parental) addresses. After (eventually) registering with a GP at their term-time address, men in particular are slow to update their addresses on the Patient Register at their new address when they move away. The extent of this

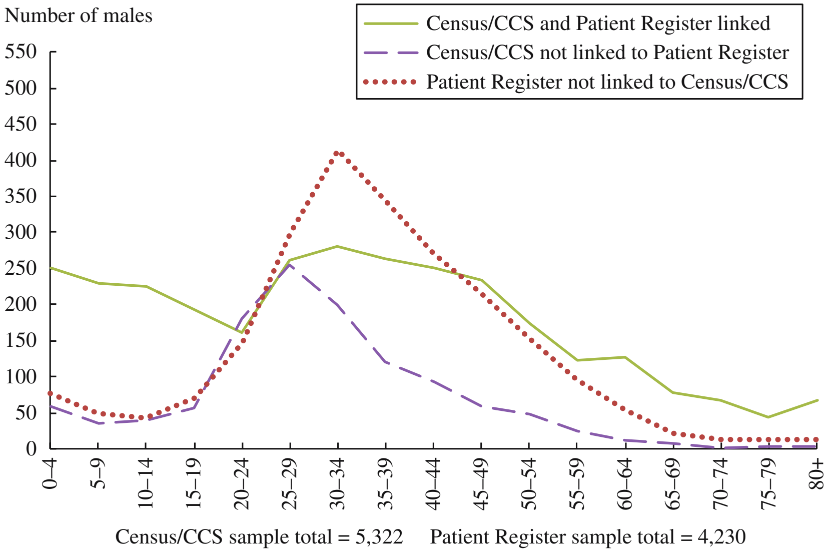

Fig. 4. Census and Patient Register linkage results for males in Inner London, by age

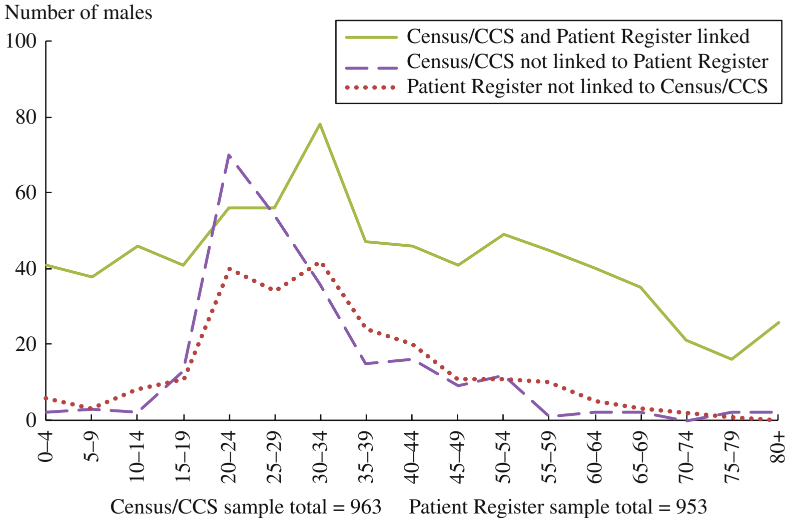

Fig. 5. Census and Patient Register linkage results for males in a university local authority, by age

disagreement between the sources cannot be deduced from the comparison of totals (Figure 6). Since the totals are similar, the comparison masks the problem that they include many people who are unique to each source.

6.4. International Migration and Excess Patient Registrations

People born outside the UK are more likely than the UK born to leave. If people do not de-register before they emigrate, their registration remains active until it gets cancelled in periodic Health Authority (HA) list-cleaning operations. The delay could cause an overcount, which contributes, for example, to the excess patient registrations seen in areas with large populations from overseas, including overseas students.

Flag 4 in the Patient Register denotes new registrations from abroad. Figure 7 compares, for each local authority, the proportion of patient registrations that did not link to the census against the proportion of unlinked records that have ‘flag 4’ status on the Patient Register. The local authorities fall into four groups:

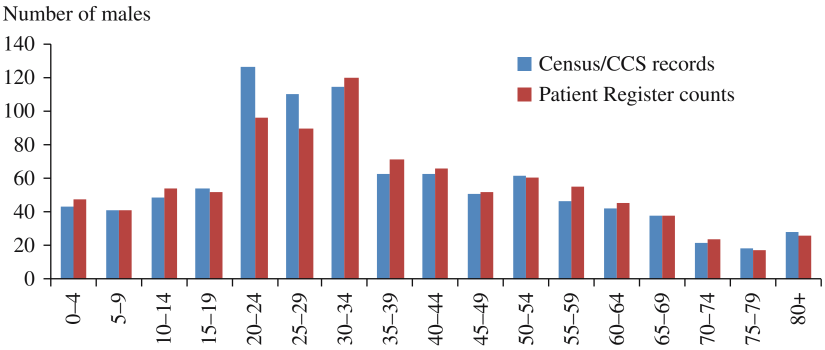

Fig. 6. Census and patient register totals for males in a university local authority, by age

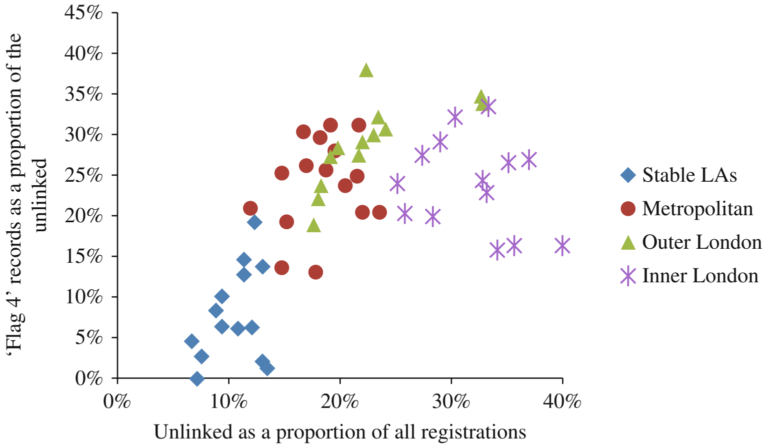

Fig. 7. Unlinked patient registrations and the proportion of international migrants within them, within local authorities. Note: **LAs with stable populations included:** *Aylesbury Vale, Blaenau Gwent, Ceredigion, Cheshire East, Knowsley, Maidstone, Mendip, Mid Devon, New Forest, Powys, Richmondshire, Tendring, Torbay.* **Metropolitan areas:** *Birmingham, Blackburn with Darwen, Bradford, Cambridge, Cardiff, Colchester, Leeds, Leicester, Liverpool, Manchester, Newcastle upon Tyne, Nottingham, Oxford, Sefton, Sheffield, Slough, Southend-on-Sea.* **Outer London:** *Barking and Dagenham, Barnet, Brent, Croydon, Ealing, Enfield, Harrow, Hounslow, Kingston upon Thames, Merton, Newham, Richmond upon Thames, Waltham Forest.* **Inner London:** *Camden, City of London, Greenwich, Hackney, Hammersmith and Fulham, Haringey, Islington, Kensington and Chelsea, Lambeth, Lewisham, Southwark, Tower Hamlets, Wandsworth, Westminster.*

- LAs with stable populations and with the least unlinked registrations, of which very few were new registrations from abroad.
- Metropolitan areas outside of London, with higher proportions of unlinked records. Among the unlinked records, between a quarter and a third were new registrations from abroad.
- Outer London local authorities, which were similar to the other metropolitan areas but tended to have higher proportions of unlinked patient registrations. (These were combined with 'Metropolitan Areas excluding Inner London' in Table 2.)
- Inner London local authorities with the highest proportions of unlinked patient registrations. Here, the proportions that were new registrations from abroad were similar to local authorities in Outer London and other metropolitan areas.

The combined evidence from the local authorities we analysed suggests that in areas with excess patient registrations, fewer than half were new registrations from abroad. Thus both internal migration and international migration are associated with excess patient registrations.

In all local authorities we found that unlinked patient register records were more likely than linked ones to be new registrations from abroad.

7. Conclusions

The 2011 Census for England and Wales used administrative data to quality assure census counts and estimates to a degree that was hitherto unprecedented and involved record linkage between census and administrative data for the first time. In the process, it revealed patterns of differential coverage in routinely collected administrative records, notably the NHS Patient Register.

The availability of rich sources of administrative data provided an unprecedented opportunity to assess and possibly enhance the quality of the census estimates, but posed some serious operational challenges. The scope for using the administrative sources was very time limited and difficult to plan ahead. The research questions to be answered by record linkage (and by extension, the administrative sources to be used) were not known in advance. In order to provide timely evidence for census quality assurance, flexibility was the key:

- Flexibility to hold and link new and upcoming sources as the census operation progressed
- Flexibility to exploit and incorporate the full range of census information as it emerged, including enumerators' 'dummy' returns for nonresponding households or derelict properties
- Flexibility to switch data linkage effort between areas as required by the Quality Assurance Panels
- Separation of person and address linkage so that linkage to the census address register would give us a head start before person-level census data were available and the quality-assurance process was fully underway
- Flexible analytical resources so that different and multiple sources could be used, as aggregates and as microdata, to address research questions that were not known in advance.

The CCS was important for the data linkage task because it provided a strategic sample that constrained the scale of record linkage and augmented the data available for analysis.

Out of 348 local authorities, we linked data in 58. These were the most challenging areas, together with some with stable populations against which we could benchmark our results. We needed high-quality linkage as the Patient Register was used by local authorities as a comparator for census estimates. We extended our linkage strategy to incorporate rules-based linkage. The use of clerical matchers was key to the success of this approach. We found inevitable discrepancies between the census and Patient Register, due to time lags in updating address information and definitional and coverage shortfalls in NHS Patient Registrations.

We found overcount in the Patient Register in areas with high levels of internal and international migration, and higher levels of overcount for men than for women. Because women visit their GPs more frequently, we speculate that they are more likely to be recorded at their current address. In some university towns, the 2011 Census data appear to be more accurate and timely for the student population than the Patient Register, as a result of a tendency for undergraduates to remain registered at their parental address. Analysis of unlinked records provided further confidence in the census estimates.

The Census Quality Assurance Panel recommended to the National Statistician that census estimates for all 348 local authorities could be published, but in the course of the quality-assurance process there were minor adjustments based on comparisons against administrative data. Even though the comparisons did not lead to any substantial amendments to census estimates, the process was very worthwhile. Firstly, it increased our confidence in the 2011 Census processes and resulting estimates; secondly, our understanding of the coverage and quality of administrative sources was greatly enhanced through both the aggregate-level comparisons and through record linkage. This provided valuable and transparent evidence to address queries that were raised about the census estimates following their publication. Thirdly, our experience of carrying out the linkage and analysis has helped to shape and inform the use of administrative data for population estimation.

7.1. Beyond 2011

In May 2010, the UK Statistics Authority asked ONS to begin a review of the future provision of population statistics in England and Wales in order to inform the government and Parliament about the options for the next census. In response, the ONS set up the Beyond 2011 Programme to undertake this work. The Programme has undertaken extensive research into and consultation on new approaches to counting the population and reviewed practices in other countries. A key focus of this work has been research into making better reuse of administrative data. The research culminated in the National Statistician making her recommendation in March 2014 (ONS 2014), which was subsequently accepted and endorsed by the Board of the UK Statistics Authority and supported by the government in July 2014.

Three key strands of work have been identified to take forward the National Statistician's recommendation:

- **2021 Census Operation** – research, development, implementation and operation of a 2021 online Census and Census Coverage Survey. At this early stage we anticipate that special attention will need to be given to online collection, the modernisation of our field processes, and making better use of administrative data.
- **Integrated Population Statistics Outputs** – integration of census, administrative and survey data to produce outputs. In this case, attention is focusing on taking forward work on data linkage, considering how administrative data can be used both to enhance the census and produce new or improved outputs.
- **Beyond 2021** – research into the shape of the census and population statistics system beyond 2021. This longer-term work will look at proposals for the future of the census and population statistics beyond 2021, including research into the potential need for new surveys after 2021 and the benchmarking of new methods.

The programme involves working with large quantities of personal information relating to everyone in England and Wales, obtained from a range of administrative sources. It is recognised that the planned approach of linking multiple administrative sources might elevate the associated risks relating to the privacy of data concerning people and households. To mitigate this risk, ONS has decided to anonymise administrative data prior to linkage to ensure that high levels of anonymity and privacy are maintained. This has resulted in developing a new method for linking anonymous data (more details are provided in ONS 2013c). Further information on the policy for safeguarding data during the research phase of the Beyond 2011 Programme can be found in ONS (2013b).

8. References

Abbott, O. 2009. “2011 UK Census Coverage Assessment and Adjustment Strategy.” *Population Trends* 137: 25–32. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/2011-uk-census-coverage-assessment-and-adjustment-methodology—article-from-population-trends-137.pdf (accessed 16 July, 2015).

Cabinet Office December. 2008. “Helping to Shape Tomorrow- the 2011 Census of Population and Housing in England and Wales.” UK: The Stationary Office.

Granquist, L. 1991. “Macro-Editing – A Review of Some Methods for Rationalising the Editing of Survey Data.” *Statistical Journal of the United Nations Economic Commission for Europe* 8: 137–154.

Li, B., H. Quan, A. Fong, and M. Lu. 2006. “Assessing Record Linkage Between Health Care and Vital Statistics Databases Using Deterministic Methods.” *BMC Health Services Research 6*. Doi: http://dx.doi.org/10.1186/1472-6963-6-48.

Li, D., S. Wang and Z. Mei. 2010. “Approximate Address Matching,” In proceedings of the International Conference on *P2P, Parallel Grid, Cloud and Internet Computing (3PGCIC)*, 4–6 November, 2010. Doi: http://dx.doi.org/10.1109/DGCIC.2010.43, 264–269, Institute of Electrical and Electronic Engineers. New York: USA.

Office for National Statistics. 2009. *2011 Census Data Quality Assurance Strategy*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/2011-census—data-quality-assurance-strategy.pdf (accessed 16 July, 2015).

Office for National Statistics. 2011a. *2011 Census – Methodology for Quality Assuring the Census Population Estimates*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/2011-census—methodology-for-quality-assuring-the-census-population-estimates.pdf (accessed 16 July, 2015).

Office for National Statistics. 2011b. *Guidance on Core to Supplementary QA*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/guidance-on-core-to-supplementary-qa.pdf (accessed 16 July, 2015).

Office for National Statistics. 2012a. *The 2011 Census Coverage Assessment and Adjustment Process*. Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/index.html (accessed 16 July, 2015).

Office for National Statistics. 2012b. *Quality Assurance of 2011 Census Population Estimates*. Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-assurance/index.html (accessed 16 July, 2015).

Office for National Statistics. 2012c. *Overcount Estimation and Adjustment*. 2011 Census: Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release–quality-assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf (accessed 16 July, 2015).

Office for National Statistics. 2012d. *Beyond 2011: Exploring the Challenges of Administrative Data*. Methods and Policies Report (M2). Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011—exploring-the-challenges-of-using-administrative-data.pdf (accessed 16 July, 2015).

Office for National Statistics. 2012e. *Comparisons Between 2011 Census Estimates and the GP NHS Patient Register Adjustment*. 2011 Census: Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/how-census-compares-with-other-data-sources/index.html (accessed 16 July, 2015).

Office for National Statistics. 2013a. *Results From Using Routinely-Collected Government Information for 2011 Census Quality Assurance*. 2011 Census: Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-assurance/index.html (accessed 16 July, 2015).

Office for National Statistics. 2013b. *Beyond 2011: Safeguarding Data for Research: Our Policy*. Methods and Policies Report (M10). Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-safeguarding-data-for-research-our-policy–m10-.pdf (accessed 16 July, 2015).

Office for National Statistics. 2013c. *Beyond 2011: Matching Anonymous Data*. Beyond 2011 M9 Methods & Policies. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-matching-anonymous-data–m9-.pdf (accessed 16 July, 2015).

Office for National Statistics. 2014. *The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority*. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation–and-recommendations/national-statisticians-recommendation.pdf (accessed 16 July, 2015).

Smallwood, S. and K. Lynch. 2010. "An Analysis of Patient Register Data in the Longitudinal Study - What Does it Tell Us About the Quality of the Data?" *Population Trends* 141: 151–169.

White, N., O. Abbott, and G. Compton. 2006. "Demographic Analysis in the UK Census: a Look Back to 2001 and Looking Forward to 2011." In Proceedings of the American Statistical Association, Survey Research Section. Alexandria, VA: American Statistical Association.

Received January 2014
Revised February 2015
Accepted March 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 475–487, http://dx.doi.org/10.1515/JOS-2015-0028

A Bayesian Approach to Population Estimation with Administrative Data

John R. Bryant[1] *and Patrick Graham*[2]

The article describes a Bayesian approach to deriving population estimates from multiple administrative data sources. Coverage rates play an important role in the approach: identifying anomalies in coverage rates is a key step in the model-building process, and data sources receive more weight within the model if their coverage rates are more consistent. Random variation in population processes and measurement processes is dealt with naturally within the model, and all outputs come with measures of uncertainty. The model is applied to the problem of estimating regional populations in New Zealand. The New Zealand example illustrates the continuing importance of coverage surveys.

Key words: Bayesian; official statistics; demography; administrative data.

1. Introduction

Statistical agencies around the world are developing new methods for population estimation that make better use of administrative data. The long-term goal is often to do away with a traditional census and to rely on administrative data, perhaps supplemented by a coverage survey. This goal has already been attained in some countries (Coleman 2013).

The conceptually simplest approach to estimating population size and structure from administrative data is to maintain a highly accurate population register, and to read population estimates straight off the register. However, few countries have this option available to them.

The conceptually simplest alternative to a population register is to take a single administrative data source, such as a list of people enrolled within the health system, and to adjust for known deficiencies. In the absence of a census, a standard way to identify deficiencies is to conduct a survey collecting information on undercoverage, overcoverage, and misclassification errors such as faulty addresses. Using a single administrative data source plus a coverage survey is much like using a traditional census plus a coverage survey. Relying on a single administrative data source has important disadvantages, however. The statistical agency is unlikely to have the same degree of control over administrative data that it does over the census, and may therefore be unable to prevent changes in policy, information technology, or recording practices that affect the

[1] Statistics New Zealand – Population Statistics, Dollan House Private Bag 4741, Christchurch 8140, New Zealand. Email: john.bryant@stats.govt.nz
[2] Statistics New Zealand, Statistical Methods, Dollan House Private Bag 4741, Christchurch 8140, New Zealand. Email: patrick.graham@stats.govt.nz

quality and consistency of the data. Moreover, a single data source may not take in all groups within the target population.

Rather than rely on a single administrative data source, a statistical agency can combine several administrative data sources. The combining of data can occur at the individual level, via record linkage. Linking together multiple administrative datasets is more difficult than is generally realised, however, particularly in countries such as New Zealand where there is no universal personal identifier. Linkage errors complicate population estimation: when an individual appears in two datasets but the individual's records are not linked, he or she may be counted twice in population estimates. Large-scale record linkage also raises privacy and ethical concerns.

The combining of data sources can instead occur at the level of the cell count. Counts classified by age, sex, and region can be calculated for each dataset, and then population can be derived as some sort of weighted combination. This avoids many of the problems of individual linking, but poses problems of its own. Assigning weights to datasets is difficult, especially when there is no gold standard and there is random variation in the data and population. Moreover, different data sources typically include different variables, and cover different age groups or time periods (Bycroft 2013; Office for National Statistics 2013).

Statistics New Zealand has been developing a formal statistical approach to deriving population estimates from multiple administrative data sources (Bryant and Graham 2013). Data are combined at the level of the cell count. The overall model contains submodels describing regularities within the demographic processes, and describing the relationship between the demographic processes and the various available datasets. The approach is Bayesian, which provides the necessary flexibility and the ability to account for diverse sources of uncertainty. Coverage rates play a central role in the modelling, as a diagnostic, and as a source of implicit weights for the data.

This article provides an overview of our approach, and describes an application to the problem of estimating regional populations in New Zealand. The application illustrates the difficulty of inferring population from administrative data alone. The results suggest that, in the absence of a traditional census, it would be necessary to supplement administrative data with a carefully-targeted coverage survey.

2. A Bayesian Framework for Population Estimation

Here we provide a brief introduction to our statistical model. More detail is available in Bryant and Graham (2013). The model is summarized in Figure 1.

At the core of the model is a demographic account Q (Rees 1979; Stone 1984). The account is a complete description of the demographic stocks and flows of interest. In Bryant and Graham (2013), the demographic account contains counts of births, deaths, migrations, and population, all disaggregated by age, sex, region, and time, and all linked by accounting identities. In the application below, however, we work with a simple account containing only population stocks. Whatever the level of detail, the account is treated as unobserved, and values for cells within the account must be inferred from available data.

Entries within an account typically exhibit strong regularities. For instance, age profiles for areas with universities typically have sharp peaks in the main student ages. The model of the demographic account, M_Q, captures these regularities. Often there are auxiliary data

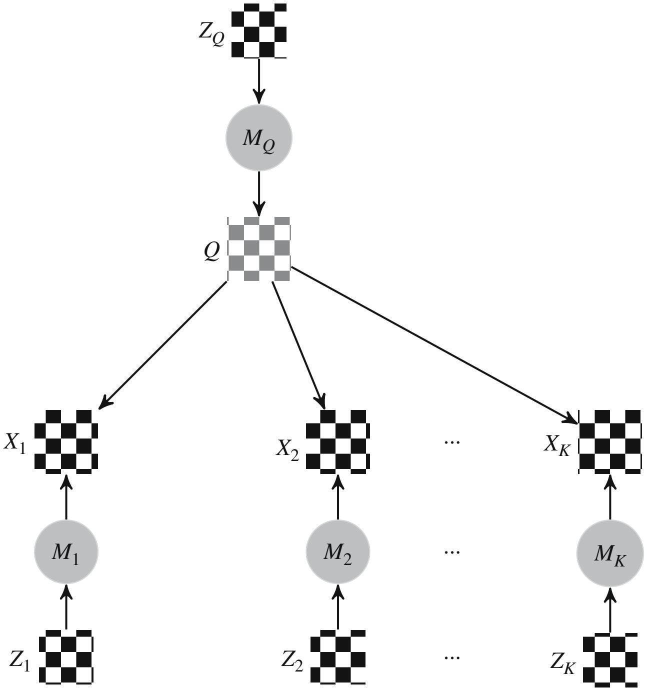

Fig. 1. Our population estimation framework. Q is the demographic account, the Xs are data sources, and the Zs are covariates. Black denotes observed quantities and grey denotes unobserved ones. Hatched squares denote counts of people or events, and circles denote submodels. Arrows denote probabilistic relationships.

Z_Q that can assist with the estimation of parameters within M_Q. Data on the location of universities, for instance, can predict the existence of age spikes.

Datasets $X_1, \ldots, X_K$ consist of counts of people or events, or proxies for these counts. No sharp distinction is made between administrative sources such as tax data and more traditional sources such as the census. Datasets can be added to or removed from the model easily.

The $M_1, \ldots, M_K$ denote data models. A data model M_k treats dataset X_k as a response and the demographic account Q as a predictor. The model describes the closeness and consistency of the relationship between the data and the underlying demographic process. The approach is similar to that of measurement error or latent variable models, in that the datasets are treated as reflecting a common unobserved construct.

The relationship between data and demographic process varies from data source to data source. With a highly reliable data source, there is essentially a one-to-one relationship. A reliable birth registration system, for instance, captures almost every birth. Some data sources are subject to undercoverage or overcoverage, but in a consistent way. For instance, a data source might cover only 80% of the target population, but maintain the same coverage level from year to year. Finally, some data sources are subject to fluctuating degrees of coverage, with no consistent relationship between coverage levels and variables such as age, sex, region, or time.

If a data source is known to be highly reliable, the data model can be designed accordingly: Section 3 gives an example. More typically, the analyst has some idea of patterns in coverage, but does not know detailed coverage rates. Data models can incorporate the analyst's qualitative knowledge by, for instance, using age and sex as predictors if the analyst thinks that coverage varies along these dimensions. The data models then provide quantitative measures of the relationship between coverage rates and the predictors. Even more can be learned when the data model is hierarchical – that is, when the coverage rates are themselves treated as draws from distributions, the parameters of which vary with the predictors (Gelman and Hill 2007). Hierarchical models can distinguish between situations where variables such as age, sex, region, and time predict coverage rates precisely and situations where predictions are poor. In other words, hierarchical models provide quantitative measures of the consistency of a data source.

When the population estimation model generates proposed values for cells in the demographic account, proposals that fit the predictions of the relevant data models are more likely to be accepted. Models for consistent data sources make sharper predictions than models for inconsistent data sources. Departures from sharp predictions are penalized more heavily than departures from diffuse ones. The population estimation model thus implicitly gives greater weight to consistent data sources than to inconsistent ones.

The fact that each dataset X_k is 'predicted' from the corresponding data model M_k and the demographic account has important practical advantages. The demographic account, by construction, has at least as much detail as any of the individual datasets. If a dataset is missing a dimension that is present in the demographic account, then the account is aggregated across that dimension before it is supplied to the data model. Similarly, if a dataset has missing values for a given year or age group, then the corresponding years or age groups are removed from the account before it is supplied to the data model. This means that it is not necessary to place all the input data into the same format. The approach thus avoids one of the most difficult and time-consuming parts of traditional population estimation.

Inference is carried out via Markov chain Monte Carlo methods. A Gibbs sampler alternates between the full conditional distributions for Q, M_Q, and $M_1, \ldots, M_k$. Sampling from the distribution for Q is difficult. The accounting identities and non-negativity constraints in Q mean that cell values do not follow standard distributions, so that customized updating procedures are required. However, sampling from M_Q and $M_1, \ldots, M_K$ is generally straightforward (Bryant and Graham 2013). The model output consists of samples from the posterior distributions for the demographic account, the demographic model M_Q, and the data models $M_1, \ldots, M_K$. Subsections 3.2 and 3.3 provide some examples.

3. Application to Subnational Population Estimation in New Zealand

3.1. *Data and Setting*

We apply a simple version of the model to the problem of estimating population counts by five-year age group, sex, time, and 'territorial authority' in New Zealand. Territorial authorities range in size from a few hundred people to 1.5 million. We omit the smallest territorial authority, and estimate counts for the remaining 66. Two of our four data sources only have consistent data for the years 2012 and 2013, so we restrict the estimation to those years. Although a population census was carried out in 2013, we do not use data from the 2013 census except for a validation exercise.

The data sources are summarized in Table 1. The first three are all administrative sources. As discussed in detail in Statistics New Zealand (2013) and Gibb (2014), information about administrative processes and comparison of counts at the national level suggest that none of the three administrative data sources accurately reflect the number of people who live in New Zealand. The target population of the primary health care data, for instance, is more or less equal to the usually resident population, but people within the target population do not appear in the data if they do not visit the doctor. Comparisons of numbers at the national level indicate that many young adults, who tend not to visit the doctor, are indeed not included. The target population for the tax data is people who have tax deducted directly from wage or social welfare payments. The target population excludes most people who do not work or receive social welfare payments, and includes some people living outside New Zealand. The target population for the electoral roll data also does not quite align with the resident population. Moreover, national figures indicate that many young people who are part of the target population are not on the electoral roll.

The one nonadministrative data source, the national-level population estimates, is the most accurate of the sources in Table 1. It is constructed by adjusting census data (in this case 2006 census data) for coverage errors, and then updated using accurate data on births, deaths, and international migration.

3.2. *Initial Model*

3.2.1. Specification

We model population using

$$q_i \sim \text{Poisson}\left(\theta_i^Q\right)$$
$$\log \theta_i^Q \sim \text{N}\left((\text{H}^Q\beta^Q)_i, \sigma_Q^2\right)$$

Table 1. Data sources used in the application

Data source	Description	Expected relationship with population counts	Detail available
Health	Enrolment in primary health care providers	Good correspondence overall, but lower for young adults, particularly males	Age, sex, region, 2012-2013
Tax	People with taxable income from work or benefits	Some overcoverage and undercoverage, varying by age and sex	Age, sex, region, 2012-2013
Electoral	People enrolled to vote	Significant undercoverage at younger ages.	Ages 18 + and region, 2013. No sex.
National population estimates	National population by age and sex	Accurate, though with some uncertainty about young adults	Age and sex, 2012-2013

where q_i is the number of people in the ith age-sex-region-time cell of the demographic account, β^Q is a vector of coefficients, and H^Q is a design matrix. The model includes age, sex, and region effects, plus all second-order interactions between these terms, plus a time effect. Priors for the model are described in the Appendix.

We model the relationship between the tax data X^{tax} and demographic account Q using

$$x_i^{\text{tax}} \sim \text{Poisson}\left(\theta_i^{\text{tax}} q_i\right)$$

$$\log \theta_i^{\text{tax}} \sim N\left((H^{\text{tax}}\beta^{\text{tax}})_i, \sigma_{\text{tax}}^2\right)$$

Parameter θ_i^{tax} measures coverage in cell i. The model includes an age effect, a sex effect, and an interaction between the two. By not including region and time effects, we are implying that we expect age-sex profiles for coverage to be similar across regions and across time. Restrictions such as this are necessary to achieve identification. The restrictions are not completely binding, however. As is apparent in the results below, a sufficiently strong signal in the data pulls the posterior distribution away from the prior.

The σ^2 term in a Bayesian hierarchical model like the one for tax measures how well the variables in β are able to explain variation in θ. A posterior distribution for σ^2 that is concentrated near zero implies that the variables have substantial predictive power (Gelman and Hill 2007). In the model for the tax data, low values for σ_{tax}^2 would imply that age, sex, and the interaction between the two accurately predict coverage rates for the tax data. In other words, low values for σ_{tax}^2would imply that the age-sex profile for coverage was approximately constant across regions and time. Conversely, high values for σ_{tax}^2 would imply inconsistent age-sex profiles.

The health and electoral data are modelled in the same way as the tax data, except that there is no sex effect in the model for the electoral data. The national population estimates need a different model. A Poisson distribution has too much variance to represent the close relationship that exists between national population estimates and the true population counts. Instead we use a Poisson-binomial mixture,

$$x_i^{\text{nat}} = u_i + v_i$$

$$u_i \sim \text{Poisson}((1 - \pi)q_i)$$

$$v_i \sim \text{Binomial}(q_i, \pi).$$

The Poisson-binomial mixture can be interpreted as a simple model of enumeration errors, in which v_i is the number of people correctly enumerated in cell i, and u_i is the number incorrectly enumerated. Parameter π is set to 0.98, based on discussions with Statistics New Zealand staff about the likely accuracy of the national estimates.

The results presented below were obtained from five independent chains with a burn-in of 10,000 and production of 10,000, recording one out of every 50 iterations. We monitored convergence by calculating potential scale reduction factors. The multivariate potential scale reduction factor (Brooks and Gelman 1998) for ten randomly chosen population cells was 1.02.

3.2.2. Results

Figure 2 shows the results for four selected regions. The first two regions are highly urban; the second two are a mix of rural areas and towns. The 95% credible intervals for the second two regions are wider, reflecting their smaller size and hence the greater relative importance of random variation. The first two regions have peaks beginning in the late teenage years, while the second two regions have troughs. These are the characteristic age profiles produced by the migration of young people out of rural areas and towns into cities.

Figure 3 shows estimates of coverage rates for the three administrative datasets. A rate of 1.0 implies that there is one person in the administrative dataset for each person in the true population; a rate higher than 1.0 implies overcoverage, and a rate lower than 1.0 implies undercoverage.

Each dataset has a characteristic age profile for coverage. The width of the credible intervals also varies across datasets. This reflects the consistency of the coverage profiles across regions and across time, or, equivalently, the value for σ. In the model for the health data, the median posterior estimate for σ is 0.013; in the model for the tax data it is 0.086; and in the model for the electoral data it is 0.055. When distributing population across regions, the model penalizes deviations from the pattern predicted by health data the most, and penalizes deviations from the pattern predicted by the tax data the least.

The results for Dunedin in Figure 3 are anomalous. The age group 20–24 appears to have coverage rates well over 1.0 in the tax and electoral data for Dunedin, but coverage rates of less than 1.0 in the tax and electoral data for other regions. The explanation for this anomaly is that the health data for Dunedin are idiosyncratic, resulting in population estimates that are too low. Dunedin has a large university and a large student population. However, the student health service in Dunedin does not belong to the standard primary health care system, so most young people there do not show up in the health data. The model has not been provided with information about the discontinuity in the relationship between health data and population. It therefore places its usual high weight on the health data and low weight on the tax and electoral data.

The tax data for the older ages shows a different sort of anomaly. Estimated coverage rises about 1.0, particularly in Auckland. The rise in apparent coverage can be explained by idiosyncrasies of the administrative data. It is clear from the metadata, and from the fact

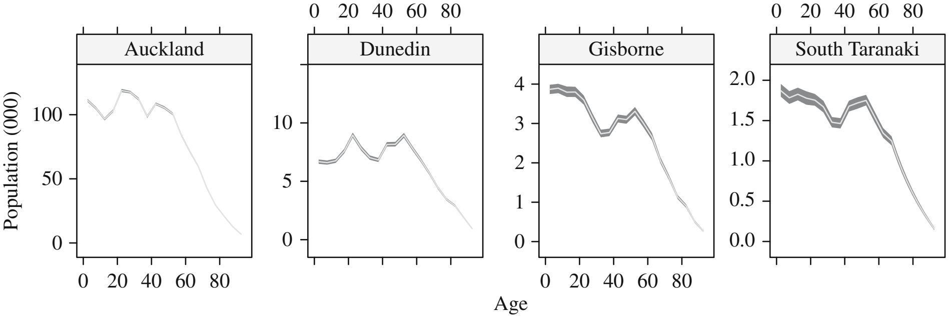

Fig. 2. Population estimates from the initial model, for four selected regions, males and females combined, 2013. The dark bands are 95% credible intervals and the grey lines are medians.

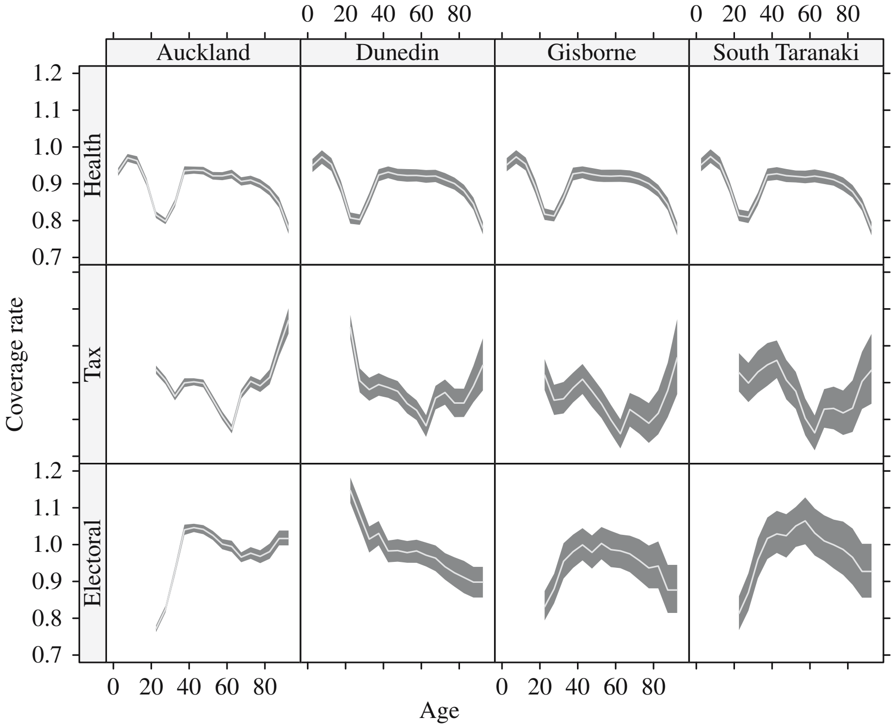

Fig. 3. Coverage rates from the initial model, for the three administrative data sources (the rows) and four selected regions (the columns), males and females combined, 2013.

that the dataset contains many people aged 100 or more, that many people are not removed from the dataset after they have died.

3.3. Revised Model

3.3.1. Specification

We make two specification changes in response to the initial results. First, we add a covariate to the population model that takes a value of 1 if a cell refers to age groups 15–19 or 20–24 and to a main centre, and 0 otherwise. This covariate captures the systematic relationship between the type of region and the number of young people. Second, we delete the cells from the health dataset that refer to 15–24 year olds in Dunedin. In the absence of data on students enrolled in the student health service, the health dataset provides little guidance on numbers of young people in Dunedin.

3.3.2. Results

Population estimates from the revised model are shown in Figure 4. The ‘student spike’ in Dunedin is substantially higher under the revised model than it was under the initial model. The credible intervals are much wider for the student ages in Dunedin than they are for other ages, which is appropriate, given that the estimates for the student ages are constructed using the two least-reliable datasets.

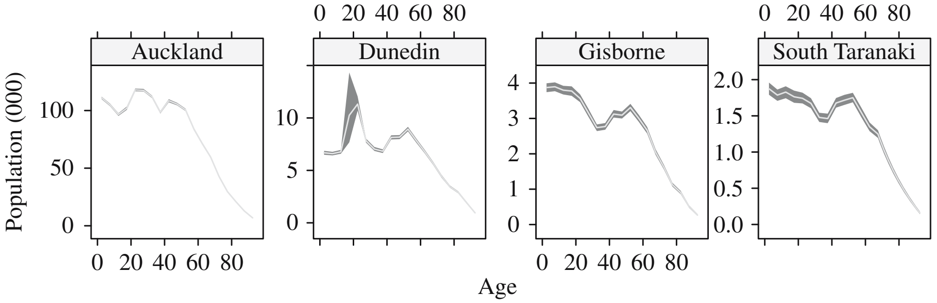

Fig. 4. Population estimates from the revised model.

Coverage rates from the revised model are shown in Figure 5. The coverage rates for the tax and electoral datasets in Dunedin look less anomalous than before, though they still differ from the other regions. We suspect that, even with the main-centre-by-age indicator variable, the population model is still pulling the Dunedin estimate down, closer to the age pattern for other regions.

Finally, Figure 6 shows results from a simple validation exercise. We take the population counts from the 2013 census and, within each age-sex combination, scale regional population numbers upwards so that they match the national population estimates described in Table 1. We subtract the model estimates from the scaled census estimates as a measure of errors in the output from the revised model. There is a clear pattern in the four

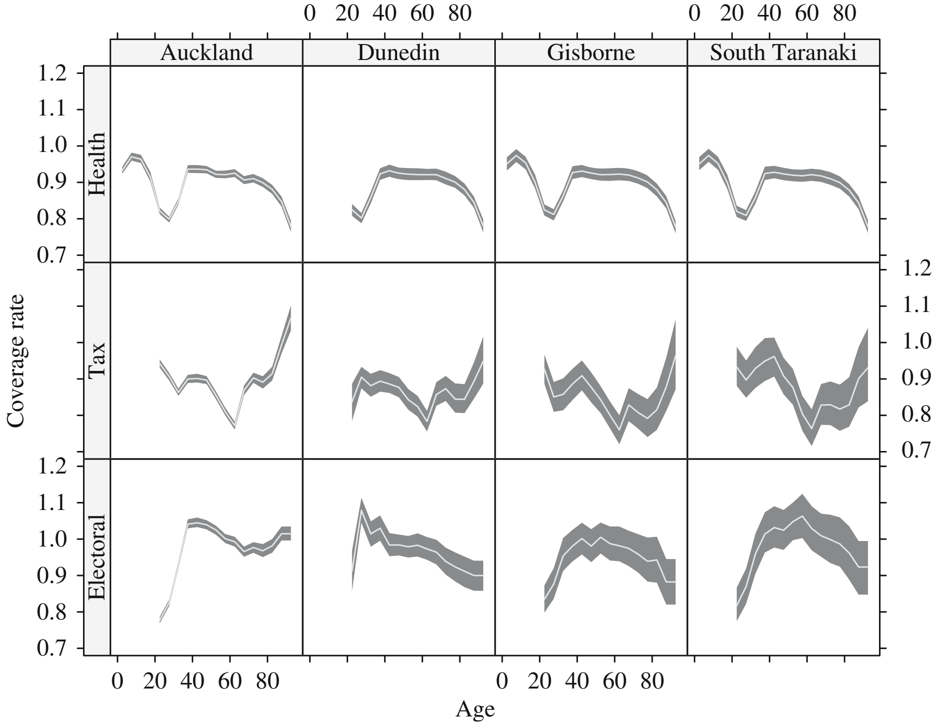

Fig. 5. Coverage rates from the revised model.

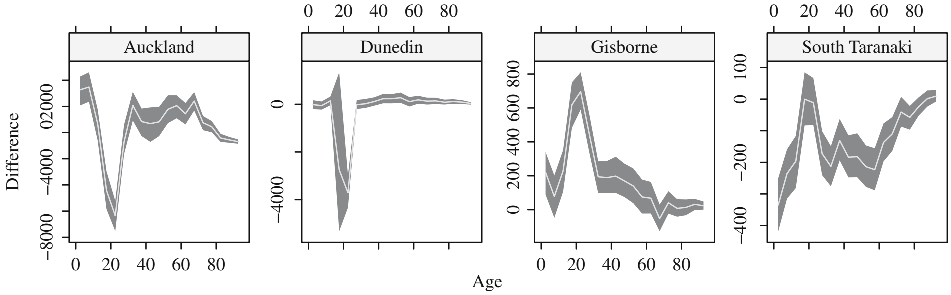

Fig. 6. Estimates from the revised model minus scaled 2013 census estimates. Values greater than 0 suggest that the model is overestimating the true population; values less than 0 suggest that it is underestimating.

selected regions (and in the remaining regions not shown here). The model consistently understates the number of young people in main centres, and overstates the number in towns and rural areas.

4. Discussion

Statistics New Zealand is developing new methods for deriving population estimates by combining counts from multiple administrative data sources. The methods implicitly weight the various data sources in proportion to the consistency of their coverage rates. The process of deriving the implicit weights is automatic and data driven. Unlike traditional approaches to the study of coverage rates, no data source needs to be treated as the gold standard. Instead, the denominator for the coverage rates is generated within the model. The methods deal naturally with random variation in the population counts and data sources. All model outputs come with measures of uncertainty.

When weighting data sources, the model does not necessarily favour data sources with higher coverage rates. For instance, if one data source has an average coverage rate of 1.0 but is inconsistent across regions or time, while another data source has an average coverage rate of 0.4 but is highly consistent, then the model weights the second dataset higher than the first. If a higher coverage rate implies greater efficiency, then data sources with higher coverage rates will tend to be more consistent. However, the distinction between high coverage and consistent coverage is important. The ability to exploit data sources with low but consistent coverage rates is an advantage of cell-level approaches to population estimation.

A typical data model in our framework simply describes the empirical relationship between the data and the demographic process, without providing reasons for any discrepancies. For instance, none of the three models for administrative data in our application distinguish between discrepancies due to misaligned target populations, discrepancies due to reporting lags, and discrepancies due to processing error, though all three types of discrepancies are present in the data. Our approach to evaluating administrative data is thus complementary to approaches such as that of Zhang (2011) which seek to identify the specific sources of error. Results from such approaches are useful for our framework as a guide to the construction of data models. In return, our approach can provide estimates of the net effect of the various errors.

The application to New Zealand regional populations presented in this article is based on a relatively simple model. Models that were used for the production of official statistics would typically be more elaborate. In particular, such models would typically be based on a full demographic account, containing births, deaths, and migrations, in order to exploit available data on these processes. The specify-estimate-evaluate cycle would be repeated many times, in the light of anomalies in coverage rates.

Nevertheless, the accuracy of any model, no matter how elaborate, is limited by the data available. If all data sources are subject to the same deficiencies, then data confrontation is unable to detect and correct for these deficiencies. An example is the overestimation of young people in rural areas and underestimation in urban areas in our modelling of regional populations in New Zealand. There is substantial evidence that administrative data systems in New Zealand miss many changes of address, or only capture them after a considerable lag (Statistics New Zealand 2013). Failure to update addresses has a greater effect on data for young people than on data for other age groups, because young people are much more mobile. The result is that administrative data for 'sending' regions contain too many young people, and administrative data for 'receiving' regions contain too few.

Such problems can be dealt with through a coverage or validation survey. The survey can be designed to respond to known problems with the administrative data. For instance, if administrative systems are failing to detect migrations by young people, then the survey can target these age groups, and ask questions about migration and the updating of addresses. The survey would yield data on true migrations versus reported migrations that could be supplied to the estimation model.

Data from the coverage survey could be included within the larger population estimation model as a special type of covariate. Models of the relationship between the coverage survey, the administrative data source, and the true population would need to include information about survey design and sample size, so that the survey data are given appropriate weight. The result would be coverage rates and population estimates that simultaneously took account of the survey data, the evidence from other data sources, and demographic plausibility.

Appendix: Further details on models

In the model for population,

$$\beta^{Q} = (\beta^{0}, \beta^{\text{age}}, \beta^{\text{sex}}, \beta^{\text{reg}}, \beta^{\text{age:sex}}, \beta^{\text{age:reg}}, \beta^{\text{sex:reg}}, \beta^{\text{time}}).$$

(For simplicity, we omit Q superscripts from the elements of β^{Q}.) Standard deviation σ_{Q} is given an improper uniform prior, as is intercept β^{0}, and the elements of sex effect β^{sex} and time effect β^{time}. The prior for the elements of age effect β^{age} is a second-order polynomial trend model, a type of dynamic linear model (Prado and West 2010, 119–120). The polynomial trend prior allows for the fact that neighbouring age groups are more likely to be similar than distant age groups. The standard deviations for the observation noise and state evolution noise in the age prior are assumed to be constant over time, and are given improper uniform priors. The elements of region effect β^{reg} are assumed to follow a Student-*t* distribution with a mean of 0 and 4 degrees of freedom.

In the initial version of the model, all interaction terms are given normal priors. The means of these priors are set to 0, and the standard deviations are given improper uniform priors. In the revised version of the model,

$$\beta_{ar}^{\text{age:reg}} \sim \text{N}(\delta + \gamma z_{ar}, \tau^2)$$

where Z_{ar} is 1 if a is age group 15–19 or 20–24 and r is "Auckland", "Christchurch", "Dunedin", "Hamilton", "Palmerston North", or "Wellington", and 0 otherwise. Parameters δ, γ, and τ are all given improper uniform priors. For identification, all subvectors within β are centered at 0 within the Gibbs sampler, with β^0 adjusted accordingly.

In the model for the tax data,

$$\beta^{\text{tax}} = (\beta^0, \beta^{\text{age}}, \beta^{\text{sex}}, \beta^{\text{age:sex}}).$$

Standard deviation σ_{tax} is given an improper uniform prior, as are intercept β^0, and the elements of sex effect β^{sex}. Age effect β^{age} is given a polynomial trend prior, identical to the prior for the age effect in the population model. Age-sex interaction $\beta^{\text{age:sex}}$ is given a normal prior with mean 0. The standard deviation for the age-sex prior is given an improper uniform prior.

The model for the health data is identical to the model for the tax data. In the model for the electoral roll data,

$$\beta^{\text{roll}} = (\beta^0, \beta^{\text{age}}).$$

Standard deviation σ_{roll} is given an improper uniform prior, as are intercept β^0, and the elements of sex effect β^{sex}. Age effect β^{age} is given a polynomial trend prior, identical to the prior for the age effect in the population model.

5. References

Brooks, S., and A. Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7: 434–455. Doi: http://dx.doi.org/10.1080/10618600.1998.10474787.

Bryant, J. R. and P. J. Graham. 2013. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis* 8: 591–622. Doi: http://dx.doi.org/10.1214/13-BA820.

Bycroft, C. 2013. *Options for Future New Zealand Censuses: Census Transformation Programme*, Technical report, Statistics New Zealand. Available at: http://www.stats.govt.nz/methods/research-papers/topss/options-future-nz-censuses.aspx (accessed 12 July, 2015).

Coleman, D. 2013. "The Twilight of the Census." *Population and Development Review* 38: 334–351. Doi: http://dx.doi.org/10.1111/j.1728-4457.2013.00568.x.

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gibb, S. 2014. *Evaluating the Potential of Linked Data Sources for Population Estimates: IDI as an Example*. Technical report, Statistics New Zealand. Available at: http://www.stats.govt.nz/methods/research-papers/topss/evaluating-potential-linked-data-sources.aspx (accessed 12 July, 2015).

Office for National Statistics. 2013. *Beyond 2011: Options Explained 2*, Technical report, Office for National Statistics. Available at: http://www.ons.gov.uk/ons/guide-method/census/2021-census/reports-library/beyond-2011-reports-archive/index.html (accessed 12 July, 2015).

Prado, R. and M. West. 2010. *Time Series: Modelling, Computation, and Inference*. New York: CRC Press.

Rees, P. 1979. "Regional Population Projection Models and Accounting Methods." *Journal of the Royal Statistical Society, Series A (General)* 142: 223–255. Doi: http://dx.doi.org/10.2307/2345082.

Statistics New Zealand 2013. *Evaluating Administrative Sources for Population Estimates.* Technical report, Statistics New Zealand. Available at: http://www.stats.govt.nz/methods/research-papers/topss.aspx#censustransformation (accessed 12 July, 2015).

Stone, R. 1984. *The Accounts of Society*, Nobel Prize in Economics documents. Available at: http://www.jstor.org/stable/2951292 (accessed 12 July, 2015).

Zhang, L.-C. 2011. "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics* 27: 415–432.

Received January 2014
Revised September 2014
Accepted November 2014

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 489–506, http://dx.doi.org/10.1515/JOS-2015-0029

Sensitivity of Mixed-Source Statistics to Classification Errors

Joep Burger[1], *Arnout van Delden*[2], *and Sander Scholtus*[2]

For policymakers and other users of official statistics, it is crucial to distinguish real differences underlying statistical outcomes from noise caused by various error sources in the statistical process. This has become more difficult as official statistics are increasingly based upon a mix of sources that typically do not involve probability sampling. In this article, we apply a resampling method to assess the sensitivity of mixed-source statistics to source-specific classification errors. Classification errors can be seen as coverage errors within a stratum. The method can be used to compare relative accuracies between strata and releases, it can assist in deciding how to optimally allocate resources in the statistical process, and it can be applied in evaluating potential estimators. A case study on short-term business statistics shows that bias occurs especially for those strata that deviate strongly from the mean value in other strata. It also suggests that shifting classification resources from small and medium-sized enterprises to large enterprises has virtually no net effect on accuracy, because the gain in precision is offset by the creation of bias. The resampling method can be extended to include other types of nonsampling error.

Key words: Accuracy; coverage error; administrative data; short-term business statistics; bootstrap; resampling.

1. Introduction

Official statistics provide information to policymakers, researchers and the general public on a country's social and economic development. Traditionally, the information is collected through sample surveys. Nowadays, National Statistical Institutes (NSIs) increasingly use administrative data. Administrative sources provide a population frame from which samples can be drawn, and auxiliary information that can be used to correct for selective nonresponse in sample surveys (Bethlehem 2009). Moreover, statistics can be based entirely on administrative data (UNECE 2007). The main advantages of administrative data are a reduced response burden and lower costs for the NSI. The costs per inhabitant of censuses based on administrative data or virtual censuses are one or two orders of magnitude smaller than those of traditional censuses (Chamberlain and Schulte Nordholt 2004), without any additional burden on respondents. On the other hand, administrative data are not designed for

[1] Statistics Netherlands, Department of Process Development and Methodology, CBS-weg 11, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. Email: j.burger@cbs.nl

[2] Statistics Netherlands, Department of Process Development and Methodology, Henri Faasdreef 312, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Emails: a.vandelden@cbs.nl and s.scholtus@cbs.nl

Acknowledgments: We thank Arjen de Boer for providing the raw data, Daniel Lewis and Piet Daas for their discussion and Bart Bakker, Peter van der Heijden, the associate editor and two anonymous referees for their useful comments on earlier drafts. This work was supported by the ESSnet on the Use of Administrative and Accounts Data for Business Statistics.

statistical purposes. They may suffer from selective undercoverage, and administrative units and variables may not match statistical definitions (Bakker and Daas 2012). In other words, they are prone to nonsampling errors along the lines of the representation side and the measurement side (Zhang 2012a). The representation side of nonsampling error addresses units, which can be redundant (out-of-scope), missing, misidentified, misclassified, and so on. The measurement side of nonsampling error addresses variables, which can be proxy, unstable, mismeasured, wrongly processed, and so forth.

To benefit from the best of both worlds, survey and administrative data can be combined at unit level through data integration techniques, such as record linkage, statistical matching, and microintegration processing. Using the strength of both sources, with the administrative data covering a large part of the population and the survey data matching statistical definitions, NSIs tend to publish statistical information at a more detailed level than with survey data alone.

It is unclear how accurate the estimates based on administrative data or mixed sources are. Knowledge of the accuracy of those estimates is crucial, both for users of the statistical output and for NSIs. For users of statistical output, statistical estimates need to be precise and approximately unbiased to achieve sound decision making. For NSIs, quantification of the accuracy can be used in the design phase of a new statistical production process to compare possible estimators and select the ‘best’ one. After the implementation of the statistical process, knowledge about the effects of various nonsampling errors on accuracy can be used to improve the production process.

The present article provides an example of the use of mixed-source estimates in business statistics. In business statistics, an important source of nonsampling errors is the classification of statistical units into economic activity or industry code. The correct industry code of a unit is hard to determine because units often perform a mixture of economic activities and their activities may change over time. For statistical purposes, the correct code can be determined using operational derivation rules and different sources, such as internet and chamber of commerce data, but finding the correct code often requires expert knowledge. NSIs often focus their editing effort on the largest and most complex units and have neither the time nor the resources to verify the industry codes for the numerous small units. Consequently, it is to be expected that some units – small units in particular – are assigned to the wrong economic activity stratum. Such classification errors can be seen as coverage errors within a stratum; a coverage error occurs when a unit is unjustly included (overcoverage) or excluded (undercoverage) from the target population. In Zhang’s (2012a) classification, these errors fall along the line of representation.

A well-developed theory for estimating the accuracy of estimates as a function of probability sampling exists that has been applied in many practical situations (e.g., Särndal et al. 1992). Far less advanced is the current theory on how to estimate the accuracy of outcomes as a function of nonsampling errors, in particular for the case of mixed sources. This theory needs to be elaborated further before it can be applied easily in practical situations. Several authors have posited ideas about this topic. Bryant and Graham (2013), for instance, proposed to estimate the uncertainty caused by nonsampling errors using a Bayesian approach. Zhang (2012b) used analytical formulas to compare the accuracy of two estimators, whereas Zhang (2011) used formulas combined with bootstrap resampling to assess uncertainty due to errors in the grouping of persons into households.

In the present article we apply a bootstrap resampling method. We limit ourselves to classification errors in business statistics, but the method can be extended to other error types and is equally applicable to social statistics. We apply the method to a case study on quarterly turnover for the short-term business statistics (STS), where data for the statistical units (enterprises) underlying the largest businesses are directly observed through a census survey and the other units are observed in administrative data. Others have considered two-phase sampling (Demnati and Rao 2009) and the case of a sample survey overlapping with a selective register (Kuijvenhoven and Scholtus 2011). We limit the results to a simple-level estimator, but the methods described can also be applied to complex estimators or to temporal changes.

The rest of the article is organized as follows. In Section 2 we develop the theory to estimate the bias and variance due to classification errors. In Section 3 we present a case study, the results of which are shown in Section 4. We close with a discussion in Section 5.

2. Theory to Estimate the Bias and Variance Due to Classification Errors

Consider a population of N units that are classified into H strata (e.g., based on economic activity). Let y_i denote the turnover – or, more generally, any quantitative variable – of unit i, and s_i the (unknown) true stratum to which this unit should be assigned. Suppose we would like to know the total turnover in each stratum: $Y_h = \sum_{i=1}^{N} a_{hi} y_i$, with

$$a_{hi} = I\{s_i = h\} = \begin{cases} 1 & \text{if } s_i = h, \\ 0 & \text{if } s_i \neq h. \end{cases}$$

In this article, we consider the relatively simple case that the true value of turnover is observed for all units. However, we do not observe the true stratum s_i but an approximation thereof, which may be affected by random classification errors. Denote the stratum to which unit i is actually assigned by $\hat{s}_i$, and let $\hat{a}_{hi} = I\{\hat{s}_i = h\}$. Then the estimated total turnover in stratum h is: $\hat{Y}_h = \sum_{i=1}^{N} \hat{a}_{hi} y_i$.

For simplicity, we suppose that random classification errors occur according to a known (or previously estimated) transition matrix $\mathbf{P} = (p_{gh})$, with $p_{gh} = \Pr(\hat{s}_i = h | s_i = g)$, where it is assumed that each unit in a given true stratum has the same probability of being misclassified in one of the other strata. (That is to say, each unit has the same transition matrix $\mathbf{P}$.) Moreover, we assume that classification errors are independent across units. Finally, we make the technical assumption that $p_{hh} > \max_{g \neq h} p_{gh}$ for all h.

In the application below, we will use a transition matrix of the following particular form:

$$\mathbf{P} = \begin{bmatrix} p & \frac{1-p}{H-1} & \cdots & \frac{1-p}{H-1} \\ \frac{1-p}{H-1} & p & \cdots & \frac{1-p}{H-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-p}{H-1} & \frac{1-p}{H-1} & \cdots & p \end{bmatrix} \qquad (1)$$

In this special case, each unit is classified correctly with probability p and misclassified with probability $1 - p$. Moreover, the misclassified units are distributed uniformly over the other strata. This simple transition matrix is used to help in the exposition of the methodology, but possible extensions are indicated in the discussion. Note that for matrices that have the form (1), the above condition $p_{hh} > \max_{g \neq h} p_{gh}$ is equivalent to $p > 1/H$.

We would like to assess the bias and variance of $\hat{Y}_h$ as an estimator for Y_h, that is,

$$B(\hat{Y}_h) = E(\hat{Y}_h - Y_h) = \sum_{i=1}^{N} \{E(\hat{a}_{hi}) - a_{hi}\} y_i, \tag{2}$$

$$V(\hat{Y}_h) = \sum_{i=1}^{N} V(\hat{a}_{hi}) y_i^2, \tag{3}$$

where in (3) we used the assumption of independent classification errors across units.

In the relatively simple situation considered here, it is not too difficult to derive analytical expressions for (2) and (3); see the Appendix for more details. Note that the resulting expressions contain unknown quantities such as Y_h that need to be estimated. Moreover, in future applications we may want to consider situations that are more complex, where this analytical approach is not possible. Therefore, this article focuses on an alternative approach to estimate (2) and (3), based on bootstrap resampling, which can be generalized to more complex situations.

For each unit i, there is an infinite population of possible classification errors, modeled by the transition probabilities $\Pr(\hat{s}_i = h | s_i = g)$ in the matrix $\mathbf{P}$. The $\hat{s}_i$ actually observed is the result of one realization of this model. Under the resampling approach, we consider a new stratum assignment variable $\hat{s}_i^*$ that is obtained by applying the transition matrix $\mathbf{P}$ to the observed $\hat{s}_i$. That is to say, we consider realisations of the alternative classification error model given by

$$\Pr(\hat{s}_i^* = h | \hat{s}_i = g) \equiv \Pr(\hat{s}_i = h | s_i = g) = p_{gh}. \tag{4}$$

We also define: $\hat{a}_{hi}^* = I\{\hat{s}_i^* = h\}$. Finally, we define the so-called bootstrap replication of the estimated total turnover in stratum h: $\hat{Y}_h^* = \sum_{i=1}^{N} \hat{a}_{hi}^* y_i$.

In terms of these bootstrap replications, the bias and variance of $\hat{Y}_h$ as an estimator for Y_h may be estimated consistently by, respectively, the bias and variance of $\hat{Y}_h^*$ as an estimator for $\hat{Y}_h$ (e.g., Efron and Tibshirani 1993). In the particular situation considered here, it is possible to obtain the latter bias and variance analytically (see the Appendix). In general, they have to be estimated through Monte Carlo simulation. For this, we generate a large number (say, R) of random draws from the classification error model (4). Denote these draws by $\hat{s}_{i1}^*, \ldots, \hat{s}_{iR}^*$. From these $\hat{s}_{ir}^*$, we can compute $\hat{a}_{hir}^* = I\{\hat{s}_{ir}^* = h\}$ and subsequently $\hat{Y}_{hr}^* = \sum_{i=1}^{N} \hat{a}_{hir}^* y_i$. The bootstrap bias and variance are then estimated as follows (Efron and Tibshirani 1993):

$$\hat{B}_R^*(\hat{Y}_h) = m_R(\hat{Y}_h^*) - \hat{Y}_h, \tag{5}$$

$$\hat{V}_R^*(\hat{Y}_h) = \frac{1}{R-1} \sum_{r=1}^{R} \{\hat{Y}_{hr}^* - m_R(\hat{Y}_h^*)\}^2, \tag{6}$$

with

$$m_R(\hat{Y}_h^*) = \frac{1}{R}\sum_{r=1}^{R} \hat{Y}_{hr}^*,$$

the average value of the bootstrap replications. For sufficiently large values of R, $\hat{B}_R^*(\hat{Y}_h)$ and $\hat{V}_R^*(\hat{Y}_h)$ converge to the true bias and variance of $\hat{Y}_h^*$ as an estimator for $\hat{Y}_h$ and hence to consistent estimators of the bias and variance of $\hat{Y}_h$.

This is an example of a parametric bootstrap method. Using the observed stratum assignments as a starting point, we resample the classification errors from an explicit model, given by the transition matrix $\mathbf{P}$. Technically, resampling model (4) can be justified as a parametric bootstrap method provided that $\hat{s}_i$ is a Maximum Likelihood Estimator (MLE) for s_i. Under the condition $p_{hh} > \max_{g \neq h} p_{gh}$ introduced above, this is indeed the case (see the Appendix).

As discussed in the Appendix, the above bootstrap estimators $\hat{B}_R^*(\hat{Y}_h)$ and $\hat{V}_R^*(\hat{Y}_h)$ are consistent but not unbiased with respect to the true bias and variance of $\hat{Y}_h$. For the special case that $\mathbf{P}$ has the form (1), it is shown in the Appendix that improved, bias-corrected bootstrap estimators may be computed as follows:

$$\hat{B}_{R,BC}^*(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1} \hat{B}_R^*(\hat{Y}_h), \tag{7}$$

$$\hat{V}_{R,BC}^*(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1} \left[\hat{V}_R^*(\hat{Y}_h) - \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K\right], \tag{8}$$

with $K = \sum_{i=1}^{N} y_i^2$. Note that, under the assumptions made here, all quantities on the right-hand sides of Expressions (7) and (8) are known. For more complex situations, analytical bias corrections for the bootstrap estimators are not readily available; we will return to this point in the discussion.

In the application below, the matrix $\mathbf{P}$ will be assumed to be known. In general, it would have to be estimated. This would require an 'audit sample' of units for which both s_i and $\hat{s}_i$ are observed. Having obtained an estimate $\hat{\mathbf{P}}$ of $\mathbf{P}$, we can apply the above bootstrap method by resampling from the classification error model (4) with $\mathbf{P}$ replaced by $\hat{\mathbf{P}}$.

3. Case Study

3.1. Data

At Statistics Netherlands, quarterly turnover for STS is based on a mix of primary and administrative data. The turnover estimates are published in four subsequent releases: 30 days, 60 days, 90 days, and one year after the end of the reference period. The turnover of most businesses is obtained from Value Added Tax (VAT) data, whereas the statistical units (enterprises) underlying the largest and most complex businesses are directly observed through a census survey. The rationale behind this design is that for larger and more complex businesses, it is not possible to make a one-to-one link between administrative units and statistical units. Furthermore, early estimates typically need to be produced before the survey and administrative data are completely available. The missing data are imputed using ratio imputation, based on data from early respondents and historical information of the nonresponding units. Because no samples are drawn and missing data are imputed, no complicated design-based or model-based estimators are required to make inferences about the target population. The estimator for the total quarterly turnover in a given industry is simply the sum of observed and imputed values over all units in both strata. More information about the case study can be found in van Delden and de Wolf (2013) and the references therein.

The turnover estimates of subsequent quarters are not only used to publish turnover growth rates – stratified by economic activity – for the STS regulation, but are also used to compute yearly turnover levels. Those turnover levels are used to calibrate results of the Structural Business Statistics (SBS), which in turn are used as one of the sources to determine the gross domestic product. Thus, for both the turnover levels and the growth rates we would like to have precise and approximately unbiased results.

We will focus on nine industries of economic activity (Figure 1), defined by the Dutch particularization of NACE Rev. 2 within Division 45: “Wholesale and retail trade and repair of motor vehicles and motorcycles”. In most of those industries, turnover estimates are based on a combination of survey and administrative data. In some industries, such as 45111 (“Import of new cars and light motor vehicles”), estimates are based mainly on survey data. In others, such as 45194 (“Wholesale and retail trade and repair of caravans”) and 45402 (“Retail trade and repair of motorcycles and related parts and accessories”), estimates are completely based on administrative data. The proportion of values that are imputed instead of observed can be substantial for early estimates (30 days after the end of the reference period) but is almost negligible for final estimates (one year after the end of the reference period).

3.2. *Parameter Values and Scenarios*

In this article, we assess the sensitivity of these estimates to classification errors. According to an internal Service Level Agreement (SLA), the three-digit NACE code should be correct for at least 95% of large enterprises (survey data) and 65% of small and medium-sized enterprises (admin data). These values resemble those of an audit held in 2000 and 2003 on the quality of the three-digit NACE code in the Dutch Business Register, which reported that 97% of the NACE codes are correct for large units (20 employees or more) in Retail Trade and 69% of the NACE codes are correct for small units (up to 19 employees) averaged over industries. The proportion of correct NACE codes is higher for large units than for small units because more resources are invested in classifying a large unit’s economic activity through profiling.

We applied the SLA figures at industry level to the survey/admin division of units, which roughly correlates with unit size. We assumed that the first two digits of the NACE code in our nine industries are correct and that the probability of moving from one industry to another is the same for all industries. We used this assumption for ease of computation, which aims to illustrate the procedure of the sensitivity analysis. Whether this assumption is valid needs to be verified by carrying out a detailed audit on classification errors within

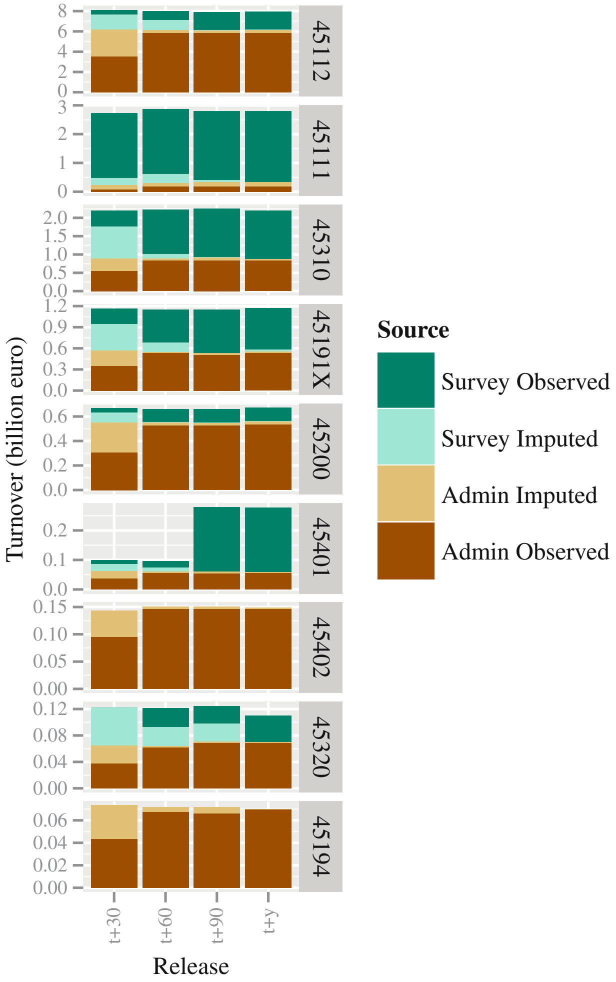

Fig. 1. Mixed-source estimates of quarterly turnover at 30 days, 60 days, 90 days and one year after the end of the reference period (third quarter of 2011) for nine industries within the Dutch particularization of NACE Rev. 2 within Division 45. Industries are ordered from large to small. Note that the y-axes are scaled independently between industries.

Division 45. The results of such an audit may lead to extensions, which are mentioned in the discussion.

We can then define two source-specific 9 × 9 transition matrices (Scenario 1):

$$\mathbf{P}^{\text{survey}} = \begin{bmatrix} \frac{19}{20} & \frac{1}{160} & \cdots & \frac{1}{160} \\ \frac{1}{160} & \frac{19}{20} & \cdots & \frac{1}{160} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{160} & \frac{1}{160} & \cdots & \frac{19}{20} \end{bmatrix}$$

and

$$\mathbf{P}^{\text{admin}} = \begin{bmatrix} \dfrac{13}{20} & \dfrac{7}{160} & \cdots & \dfrac{7}{160} \\ \dfrac{7}{160} & \dfrac{13}{20} & \cdots & \dfrac{7}{160} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{7}{160} & \dfrac{7}{160} & \cdots & \dfrac{13}{20} \end{bmatrix}$$

Note that both matrices are special cases of the matrix **P** in (1).

Although it makes intuitive sense to allocate more resources to large units that have a large impact on the statistical outcome, one could also argue that many small units may still have a considerable impact and should not be ignored altogether. In order to study the relative importance of resource allocation, we introduce a second scenario. By switching the matrices between sources, we studied what would happen if instead 65% of large enterprises (survey data) and 95% of small and medium-sized enterprises (admin data) were correctly classified for economic activity (Scenario 2). In summary, we are comparing a scenario where classification resources are mainly allocated to large units receiving a questionnaire with a scenario where classification resources are mainly allocated to small units whose information is derived from administrative sources.

3.3. Resampling

Using this input, we first drew a new industry code for each unit from these transition matrices. For instance, a unit that receives a survey and is classified in industry 45111 has a probability of 19/20 of remaining in 45111 and a probability of 1/160 of ending up in one of the other eight industries. A unit for which the data come from the admin source and that is classified in industry 45111 has a probability of 13/20 of remaining in 45111 and a probability of 7/160 of ending up in one of the other eight industries. We then recalculated the population parameter per (new) industry. Next, we repeated this a large number of times: $R = 10,000$ simulations per estimate, which seemed sufficient for confidence intervals to converge (Burger et al. 2013). From these replications, the bias and variance due to classification errors were estimated using the bias-corrected expressions (7) and (8). In summary, we assumed a stochastic error process and we used resampling to quantify the effects of this error process on the turnover estimates.

4. Results

Each turnover estimate is compared with the distribution of bootstrap replications in Figure 2a. The estimated variance and the square of the bias were added together, resulting in the mean square error (MSE) as a measure of accuracy. The square root (RMSE) was taken to revert to the unit of the data (euro), and was normalized (relative root mean squared error; RRMSE) to the total turnover estimated from observed and imputed data to make estimates comparable between releases and industries (Figure 2b).

The RRMSE can be alarmingly high: over 900% (Figure 2). We would like to stress, however, that we have estimated not the true accuracy of the turnover estimates, but their

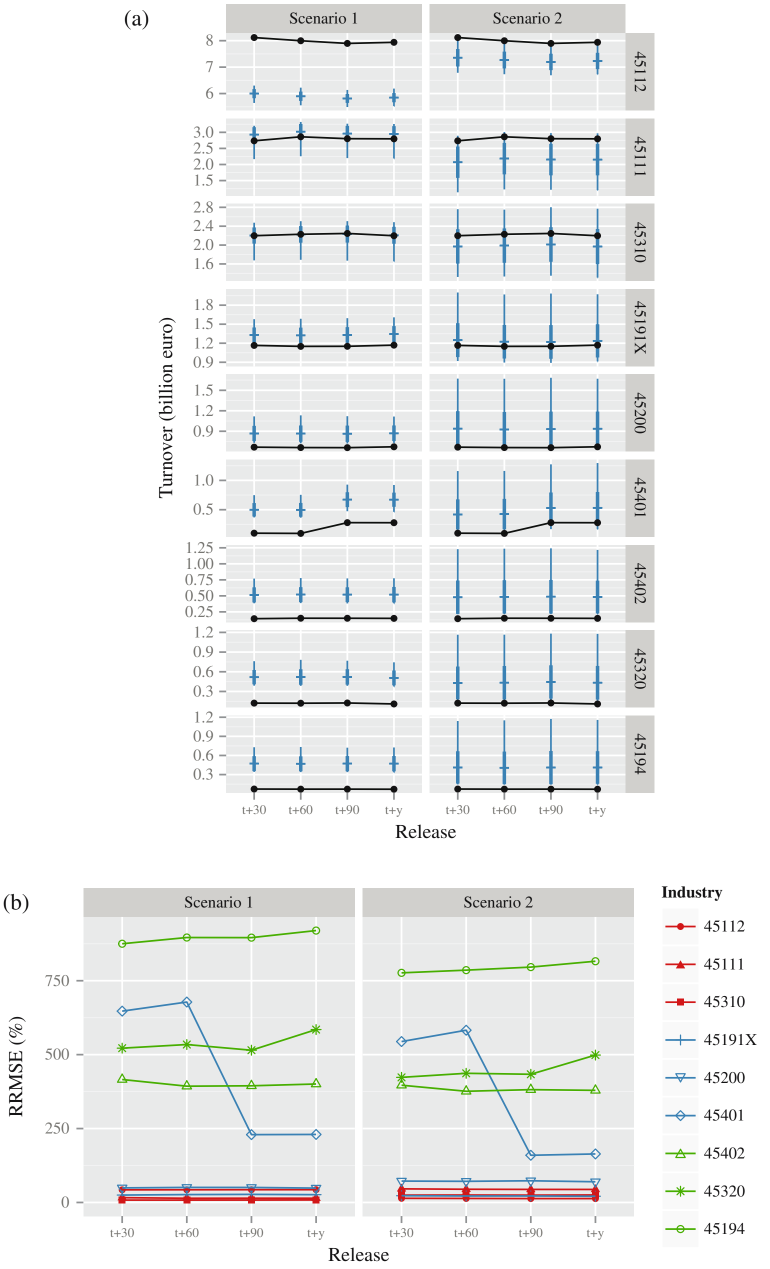

Fig. 2. Sensitivity of mixed-source estimates to source-specific classification error. (a) Quarterly turnover per industry and release estimated from observed and imputed data (black dots and lines), and simulated mean (blue horizontal dashes) ± SD (blue thick bars), and 2.5th and 97.5th percentiles (blue thin bars) using 10,000 simulations per estimate. Note that the y-axes are scaled independently between industries. (b) Root mean square error normalized to the quarterly turnover estimated from observed and imputed data. Classification error is assumed largest in admin stratum (Scenario 1) or survey stratum (Scenario 2). Industries are ordered from large to small.

relative sensitivity to classification errors. In particular, having uniform transition probabilities between strata may not be a realistic assumption. Moreover, the RRMSE correlates negatively with the turnover estimates, that is, only small industries (a few hundred million euros or less) have such a high RRMSE.

Simulations under Scenario 1 show that source-specific misclassification can result in strongly biased estimates (Figure 2). Our dataset contains one high-turnover industry (45112). In Figure 2, the simulated total turnover for this industry lies consistently below the original estimate. According to Expressions (5) and (7), this means that the total turnover of this industry is underestimated relative to the unknown true value. This bias may be explained as follows. First, the turnover in 45112 is substantially based on units using the admin data (Figure 1), which have a fair chance of being misclassified. Second, misclassified units from other industries that are classified erroneously in 45112 typically have low turnover. Similarly, the total turnover of low-turnover industries such as 45194 is overestimated relative to the unknown true value, because many small units are erroneously replaced by units from higher-turnover industries. This confirms the analytical solution showing that the absolute bias increases the more the turnover of an industry deviates from the average turnover of the other industries (see the Appendix). In industry 45401, late estimates are more accurate than early estimates because they are based on more units with a likely correct industry code (survey data, see also Figure 1). In the other industries we do not observe an effect of release on accuracy because the ratio between survey and administrative data remains fairly constant and the imputed values were held fixed (see the discussion).

When we assume that the economic activity is more reliable for small and medium-sized enterprises than for large enterprises (Scenario 2), our estimates are indeed less precise, but also less biased (Figure 2). This suggests that shifting the focus of editing the industry classification from small and medium-sized enterprises to large enterprises can result in more biased estimates. Such a shift in resources has virtually no net effect on accuracy of the level estimates (see Figure 2b), because the gain in precision is offset by the creation of bias.

For the simple scenarios used here, it is possible to derive analytical expressions for the bias-corrected bootstrap estimators of bias and variance; see Expressions (17) and (18) in the Appendix. Note that we can apply these expressions separately to survey and admin data, as there is no interaction between the two data sources in this study. For Scenario 1, working out Expressions (17) and (18) with $H = 9$ and $p = \frac{19}{20}$ (survey data) or $p = \frac{13}{20}$ (admin data), we find:

$$\hat{B}^*_{\infty,BC}(\hat{Y}_h) = \frac{8}{151}\left\{\bar{\hat{Y}}^{(-h),\text{survey}} - \hat{Y}_h^{\text{survey}}\right\} + \frac{56}{97}\left\{\bar{\hat{Y}}^{(-h),\text{admin}} - \hat{Y}_h^{\text{admin}}\right\},$$

and

$$\hat{V}^*_{\infty,BC}(\hat{Y}_h) = \frac{38}{755}\hat{K}_h^{\text{survey}} + \frac{159}{24160}\sum_{g\neq h}\hat{K}_g^{\text{survey}} - \frac{311}{483200}K^{\text{survey}} + \frac{182}{485}\hat{K}_h^{\text{admin}} + \frac{1071}{15520}\sum_{g\neq h}\hat{K}_g^{\text{admin}} - \frac{12593}{310400}K^{\text{admin}}.$$

In these expressions, $\hat{Y}_h^X = \sum \hat{a}_{hi} y_i$, $\hat{K}_h^X = \sum \hat{a}_{hi} y_i^2$, and $K^{\mathrm{X}} = \sum y_i^2$ where the sums are over all units in source X, with $X \in \{\text{survey}, \text{admin}\}$; in addition, $\overline{\hat{Y}}^{(-h),\mathrm{X}} = \frac{1}{H-1} \sum_{g \neq h} \hat{Y}_g^X$. Analogous expressions are obtained for Scenario 2 by interchanging the coefficients for survey data and admin data.

The numerical solution for the bias and standard deviation closely resembles the analytical solution derived in the Appendix (Figure 3). The mean difference in bias between the numerical and analytical solution is zero euro with the maximum absolute difference being merely twelve million euros (eleven percent of the analytical solution). The mean relative difference in standard deviation is 0.6% with the maximum relative absolute difference being 7.4% of the analytical solution. This confirms that 10,000 simulations are sufficient to approximate the analytical solution.

5. Discussion

For policymakers and other users of official statistics, it is crucial to distinguish real differences between statistical outcomes from noise caused by various error sources in the statistical process. This has become more difficult as official statistics are now increasingly based upon a mix of sources that typically do not involve probability sampling. We have described a case study where statistical units (enterprises) underlying large and complex businesses are directly observed through a census survey and the turnover of smaller and less complex enterprises is obtained from tax data.

The resampling method described in the current article provides insight into the sensitivity of mixed-source statistics to a source-specific nonsampling error. Results can be used to compare industries and releases, and can assist in deciding where to invest resources into the statistical process. Our results show that bias occurs especially in those strata that deviate strongly from the mean value in other strata. The example we have shown also suggests that shifting classification resources from small and medium-sized enterprises to large enterprises has virtually no net effect on the accuracy of the level estimates, because the gain in precision is offset by the creation of bias. On the other hand, this resource allocation might improve the accuracy of temporal turnover changes, because the creation of bias in both time points is annihilated, whereas the gain in precision is not. Results indicate that level estimates will become less biased when NSIs

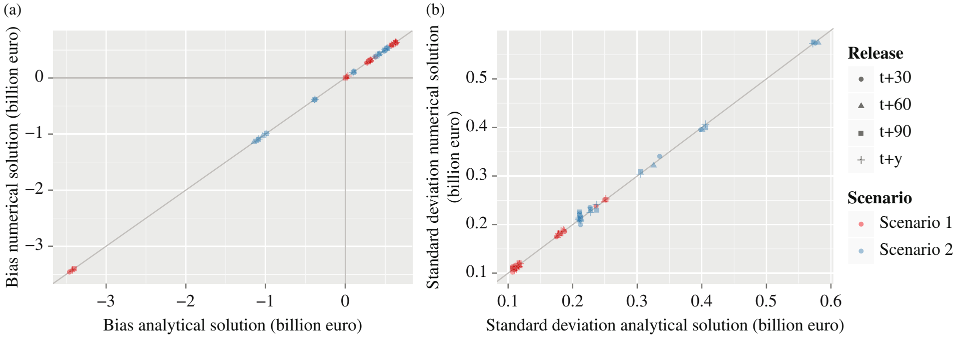

Fig. 3. Comparison between analytical and numerical solution for (a) bias and (b) standard deviation. Diagonal shows y = x.

find ways to improve the correctness of the industry codes of small enterprises, while maintaining the industry code quality of large enterprises. Because manual coding will be too expensive in practice, other approaches are needed. One possible future direction is to automatically collect data on products and services from business websites combined with text-mining techniques to translate the results into reliable industry codes.

The resampling method that we have presented can be used not only for sensitivity analyses but also to estimate the accuracy of outcomes. A major prerequisite to achieving this is to find cost-effective ways that can be used by NSIs to obtain a sound estimation of the error distribution. In our case study, we used reasonable parameter values for the probability that the observed industry code is correct. Nonetheless, we simply assumed that the probability of moving from one industry to another is the same, whereas in reality we expect those probabilities to vary, both between pairs of strata and between units. With our current parameter settings, we found extremely high RRMSEs in some industries. These results underline that our parameterization needs to be refined before drawing final conclusions about the data. We encourage other NSIs to run similar simulations with their own parameter settings of the transition matrix.

We see two steps to improve estimating a transition matrix. First, we need to understand which variables determine the correctness of an industry code for a specific unit, for instance its (observed) size class, its three-digit NACE code and the occurrence of an event (birth, merger, take-over etc.). Second, we need to estimate the error distribution. Possibilities for estimating the desired input would be to compare different sources, to derive estimates from the editing process, to apply audit sampling, and/or to model the true economic activity as a latent class. Note that accuracy estimates can also be extended to account for uncertainty in knowledge about those parameter values. Zhang (2011) used bootstrap resampling to account for that issue. In a Bayesian approach, uncertainty about the parameter values would be modeled by a prior distribution.

NSIs typically develop new estimators as new data sources become available or the statistical process is redesigned. The resampling method can also be applied to compare different estimators and to test which estimator is the least sensitive to the error process. It could also be used to decide about the line of demarcation between the survey and the admin data.

Note that we have assumed that the imputed turnover values are independent of the industry code. In reality, the industry code is used as auxiliary information in the imputation process. It would therefore be more realistic to impute missing values after resampling instead of assuming fixed imputed values (Shao and Sitter 1996). This would affect early releases where a substantial proportion of the estimate is based on imputed values. We expect that, when variation due to imputation is accounted for, classification errors will affect early releases more than late ones.

A theoretical difficulty that remains to be solved is that the direct bootstrap estimators of bias and variance may be biased in practice. In the above simplified application, we could correct this bias analytically. However, we also want to be able to use the bootstrap method in more realistic situations (as discussed above) where analytical derivations are no longer feasible, and we have no reason to assume that the bootstrap estimators will be less biased in these applications. It may be possible to obtain bias corrections to the bootstrap estimators numerically, for example, by applying a nested version of the bootstrap in which the bootstrap resamples are resampled themselves; see Efron and Tibshirani (1993). Another, computationally more attractive possibility could be to work with so-called bias-corrected bootstrap confidence intervals (Efron and Tibshirani 1993; DiCiccio and Efron 1996) instead of bias and variance estimates. This remains to be investigated.

The resampling method could be adapted to specific situations or needs. First of all, we could extend the method to account for overcoverage and undercoverage of units in the population frame. To that end, we could introduce an exclusion stratum, ‘outside the population’, and for each industry code estimate the overcoverage (true value is ‘outside the population’) and the undercoverage: the proportion that is unjustly missing. Furthermore, we could extend the method to study measurement errors, a combination of (interacting) nonsampling errors or errors due to nonprobability sampling (see for instance de Munnik et al. 2013). Another extension could be to assess the effect on accuracy of changes over time rather than of levels.

Appendix

The Observed Industry Code As An MLE for the True Industry Code

Recall from Section 2 that the resampling model (4) can be justified as a parametric bootstrap method provided that $\hat{s}_i$ is a Maximum Likelihood Estimator (MLE) for s_i. Below we will prove that this is the case.

Let $\boldsymbol{s} = (s_1, \ldots, s_N)'$ and $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_N)'$ denote vectors of true and observed industry codes, respectively. Since classification errors are assumed to be independent across units, the joint parametric model for the observed industry codes is given by:

$$\Pr\left(\hat{\boldsymbol{s}} = (h_1, \ldots, h_N)' | \boldsymbol{s} = (g_1, \ldots, g_N)'\right) = \prod_{i=1}^{N} \Pr\left(\hat{s}_i = h_i | s_i = g_i\right) = \prod_{i=1}^{N} p_{g_i h_i}.$$

Consider the log-likelihood function of the unknown parameter vector s, given the observed industry codes $\hat{s}$. By definition, it holds that:

$$\log L\left(\boldsymbol{s} = (g_1, \ldots, g_N)' | \hat{\boldsymbol{s}} = (h_1, \ldots, h_N)'\right) = \sum_{i=1}^{N} \log p_{g_i h_i}.$$

Since we assumed independence across units, we can maximize this sum by maximizing each term separately. Under the condition that $p_{hh} > \max_{g \neq h} p_{gh}$ for all h, it follows that the i^{th} term is maximized by choosing $s_i = g_i = h_i = \hat{s}_i$. We conclude that the MLE of $\boldsymbol{s}$ is given by $\hat{\boldsymbol{s}}$. As noted in Section 2, this justifies the use of resampling model (4) as an application of the parametric bootstrap. In addition, it follows that $\hat{Y}_h$ is a so-called ‘plug-in estimator’ of Y_h, which justifies Expression (5) (Efron and Tibshirani 1993).

While $\hat{s}_i$ is the MLE of s_i here, it will be shown below that the direct bootstrap estimators (5) and (6) are biased with respect to (2) and (3). This may be explained by the fact that we are using a sample of size N to estimate the N unknown parameters $s_1, \ldots, s_N$ of the parametric model. It is well known that MLEs – and, by extension, bootstrap estimators – are usually biased in situations where the effective sample size is small.

On the other hand, the bootstrap estimators *are* asymptotically consistent, because we are *not* in a situation where the number of unknown parameters increases with the sample size. Given our fixed population of N units, we could – in theory – obtain m independently assigned industry codes $\hat{s}_{i1}, \ldots, \hat{s}_{im}$ for each unit, thereby drawing a sample of size mN from the parametric model. The bias in the corresponding bootstrap estimators – with Model (4) applied to the MLE of s_i based on $\hat{s}_{i1}, \ldots, \hat{s}_{im}$ – would then vanish as $m \longrightarrow \infty$.

Derivation of Bias and Variance

For the highly simplified situation considered in Section 2, we can derive analytical expressions for the bias and variance of $\hat{Y}_h$.

Let $\boldsymbol{a}_i = \left(a_{1i}, \ldots, a_{Hi}\right)'$ and $\hat{\boldsymbol{a}}_i = (\hat{a}_{1i}, \ldots, \hat{a}_{Hi})'$. Given that classification errors are described by a transition matrix $\mathbf{P} = (p_{gh})$, we observe that:

$$E(\hat{a}_{hi}) = \sum_{g=1}^{H} a_{gi} E\left(\hat{a}_{hi} | s_i = g\right) = \sum_{g=1}^{H} a_{gi} \Pr\left(\hat{s}_i = h | s_i = g\right) = \sum_{g=1}^{H} a_{gi} p_{gh},$$

and hence that $E(\hat{\boldsymbol{a}}_i) = \mathbf{P}'\boldsymbol{a}_i$. Here we used that $a_{gi} = 1$ for exactly one $g \in \{1, \ldots, H\}$. Now let $\boldsymbol{y} = \left(Y_1, \ldots, Y_H\right)'$ and $\hat{\boldsymbol{y}} = (\hat{Y}_1, \ldots, \hat{Y}_H)'$ denote vectors of (estimated) stratum totals. By definition, $\boldsymbol{y} = \sum_{i=1}^{N} \boldsymbol{a}_i y_i$ and $\hat{\boldsymbol{y}} = \sum_{i=1}^{N} \hat{\boldsymbol{a}}_i y_i$. Noting that $E\left(\hat{\boldsymbol{y}}\right) = \sum_{i=1}^{N} E(\hat{\boldsymbol{a}}_i) y_i = \mathbf{P}'\boldsymbol{y}$, we obtain for the bias of $\hat{\boldsymbol{y}}$:

$$B\left(\hat{\boldsymbol{y}}\right) = E\left(\hat{\boldsymbol{y}}\right) - \boldsymbol{y} = \left(\mathbf{P}' - \mathbf{I}\right)\boldsymbol{y}, \tag{9}$$

with $\mathbf{I}$ denoting the $H \times H$ identity matrix. In particular, this yields the following expression for the bias of a single stratum total (2):

$$B\left(\hat{Y}_h\right) = \left(p_{hh} - 1\right) Y_h + \sum_{g \neq h} p_{gh} Y_g.$$

In the special case that $\mathbf{P}$ has the Form (1), this expression can be simplified to:

$$B\left(\hat{Y}_h\right) = \left(p - 1\right) Y_h + \frac{1 - p}{H - 1} \sum_{g \neq h} Y_g = \left(1 - p\right)\left\{\bar{Y}^{(-h)} - Y_h\right\}, \tag{10}$$

where $\bar{Y}^{(-h)} = \frac{1}{H-1} \sum_{g \neq h} Y_g$ is the average stratum total over all strata *except* stratum h. This formula shows that the (absolute) bias decreases with p, as expected. It also shows that the (absolute) bias increases the further Y_h deviates from $\bar{Y}^{(-h)}$. In other words, bias occurs especially for those strata that deviate strongly from the mean value in other strata.

Next, we consider the variance of $\hat{\boldsymbol{y}}$. Since $\hat{\boldsymbol{a}}_i$ contains binary values, it holds that $\hat{\boldsymbol{a}}_i \hat{\boldsymbol{a}}_i{}' = \text{diag}(\hat{\boldsymbol{a}}_i)$, where diag($\boldsymbol{x}$) denotes the diagonal matrix with $\boldsymbol{x}$ on the main diagonal. Similarly, $\boldsymbol{a}_i \boldsymbol{a}_i{}' = \text{diag}(\boldsymbol{a}_i)$. Therefore, the variance-covariance matrix of $\hat{\boldsymbol{a}}_i$ may be written as follows:

$$V(\hat{\boldsymbol{a}}_i) = E(\hat{\boldsymbol{a}}_i \hat{\boldsymbol{a}}_i{}') - E(\hat{\boldsymbol{a}}_i) E(\hat{\boldsymbol{a}}_i{}') = \text{diag}(E(\hat{\boldsymbol{a}}_i)) - \mathbf{P}'\boldsymbol{a}_i \boldsymbol{a}_i{}'\mathbf{P} = \text{diag}(\mathbf{P}'\boldsymbol{a}_i) - \mathbf{P}'\text{diag}(\boldsymbol{a}_i)\mathbf{P},$$

where we used $E(\hat{\boldsymbol{a}}_i) = \mathbf{P}'\boldsymbol{a}_i$ as derived above. Now using the fact that the variance-covariance matrix $V(\hat{\boldsymbol{y}})$ can be written as $V(\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} V(\hat{\boldsymbol{a}}_i) y_i^2$ [cf. Expression (3)], we obtain:

$$V(\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} \left\{ \mathrm{diag}(\mathbf{P}'\boldsymbol{a}_i y_i^2) - \mathbf{P}'\mathrm{diag}(\boldsymbol{a}_i y_i^2)\mathbf{P} \right\} = \mathrm{diag}(\mathbf{P}'\boldsymbol{k}) - \mathbf{P}'\mathrm{diag}(\boldsymbol{k})\mathbf{P}. \qquad (11)$$

Here, $\boldsymbol{k} = \left(K_1, \ldots, K_H\right)'$, with K_h denoting the sum of squared values for variable y_i in stratum h; that is, $K_h = \sum_{i=1}^{N} a_{hi} y_i^2$ and $\boldsymbol{k} = \sum_{i=1}^{N} \boldsymbol{a}_i y_i^2$. In particular, the main diagonal of $V\left(\hat{\boldsymbol{y}}\right)$ contains the following elements:

$$V\left(\hat{Y}_h\right) = \sum_{g=1}^{H} p_{gh} K_g - \sum_{g=1}^{H} p_{gh}^2 K_g = \sum_{g=1}^{H} p_{gh}(1 - p_{gh}) K_g.$$

In the special case that **P** has the Form (1), this formula simplifies to:

$$V(\hat{Y}_h) = p\left(1 - p\right) K_h + \frac{1-p}{H-1}\left(1 - \frac{1-p}{H-1}\right) \sum_{g \neq h} K_g. \qquad (12)$$

Application to the Bootstrap Estimators and Derivation of (7) and (8)

Since the bootstrap replications $\hat{Y}_h^*$ are obtained by resampling from the classification error model (4), analogous analytical expressions to (9) and (11) may be derived for the bias and variance-covariance matrix of the bootstrap replications: $B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = (\mathbf{P}' - \mathbf{I})\hat{\boldsymbol{y}}$ and $V(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = \mathrm{diag}(\mathbf{P}'\hat{\boldsymbol{k}}) - \mathbf{P}'\mathrm{diag}(\hat{\boldsymbol{k}})\mathbf{P}$. Thus, for the case study in Section 3, it was possible to obtain bootstrap estimates of the bias and variance of the original estimators *without* resorting to Monte Carlo simulations. We denote these analytical estimates by $\hat{B}_\infty^*\left(\hat{Y}_h\right)$ and $\hat{V}_\infty^*\left(\hat{Y}_h\right)$, to indicate that the same estimates would also be obtained by taking the limit $R \rightarrow \infty$ in (5) and (6). In particular, for the special case that **P** has the Form (1), we obtain [cf. (10) and (12)]:

$$\hat{B}_\infty^*\left(\hat{Y}_h\right) = \left(1 - p\right)\left\{\overline{\hat{Y}}^{(-h)} - \hat{Y}_h\right\}, \qquad (13)$$

$$\hat{V}_\infty^*\left(\hat{Y}_h\right) = p\left(1 - p\right)\hat{K}_h + \frac{1-p}{H-1}\left(1 - \frac{1-p}{H-1}\right) \sum_{g \neq h} \hat{K}_g, \qquad (14)$$

in obvious notation.

It is not difficult to show that the above bootstrap estimators are biased with respect to the true bias and variance of $\hat{Y}_h$. In fact, we have:

$$E\{\hat{B}_\infty^*\left(\hat{\boldsymbol{y}}\right)\} = (\mathbf{P}' - \mathbf{I})E(\hat{\boldsymbol{y}}) = (\mathbf{P}' - \mathbf{I})\mathbf{P}'\boldsymbol{y} = \mathbf{P}'(\mathbf{P}' - \mathbf{I})\boldsymbol{y} = \mathbf{P}'B(\hat{\boldsymbol{y}})$$

according to Expression (9). Similarly,

$$\begin{aligned}E\left\{\hat{V}^*_\infty\left(\hat{\boldsymbol{y}}\right)\right\} &= \mathrm{diag}(\mathbf{P}'E(\hat{\boldsymbol{k}})) - \mathbf{P}'\mathrm{diag}(E(\hat{\boldsymbol{k}}))\mathbf{P} \\ &= \mathrm{diag}(\mathbf{P}'B(\hat{\boldsymbol{k}})) + \mathrm{diag}(\mathbf{P}'\boldsymbol{k}) - \mathbf{P}'\mathrm{diag}(B(\hat{\boldsymbol{k}}))\mathbf{P} - \mathbf{P}'\mathrm{diag}(\boldsymbol{k})\mathbf{P} \\ &= V\left(\hat{\boldsymbol{y}}\right) + \mathrm{diag}\left(\mathbf{P}'\left(\mathbf{P}' - \mathbf{I}\right)\boldsymbol{k}\right) - \mathbf{P}'\mathrm{diag}((\mathbf{P}' - \mathbf{I})\boldsymbol{k})\mathbf{P}.\end{aligned}$$

In the last line, we used Expression (11). We also used the fact that $B\left(\hat{\boldsymbol{k}}\right) = \left(\mathbf{P}' - \mathbf{I}\right)\boldsymbol{k}$, by analogy with Expression (9). This shows that, in the presence of classification errors, $E\left\{\hat{B}^*_\infty\left(\hat{\boldsymbol{y}}\right)\right\} \neq B\left(\hat{\boldsymbol{y}}\right)$ and $E\left\{\hat{V}^*_\infty\left(\hat{\boldsymbol{y}}\right)\right\} \neq V\left(\hat{\boldsymbol{y}}\right)$.

For the special case that **P** has the Form (1), we can simplify the above expression for $E\left\{\hat{B}^*_\infty\left(\hat{\boldsymbol{y}}\right)\right\}$ to:

$$E\left\{\hat{B}^*_\infty\left(\hat{Y}_h\right)\right\} = pB\left(\hat{Y}_h\right) + \frac{1-p}{H-1}\sum_{g\neq h} B\left(\hat{Y}_g\right) = \left(p - \frac{1-p}{H-1}\right)B\left(\hat{Y}_h\right). \qquad (15)$$

Here, we used the fact that the overall total turnover $Y = \sum_{h=1}^H Y_h = \sum_{i=1}^N y_i$ is not affected by classification errors; hence, $\sum_{h=1}^H \hat{Y}_h = Y$ and $\sum_{g\neq h} B(\hat{Y}_g) = -B(\hat{Y}_h)$. A similar, slightly more tedious derivation shows that, in this special case:

$$E\{\hat{V}^*_\infty(\hat{Y}_h)\} = \left(p - \frac{1-p}{H-1}\right)V(\hat{Y}_h) + \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K, \qquad (16)$$

with $K = \sum_{h=1}^H K_h = \sum_{i=1}^N y_i^2$.

To derive the bias-corrected bootstrap estimators (7) and (8), we rearrange Expressions (15) and (16) as follows:

$$B(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1} E\{\hat{B}^*_\infty(\hat{Y}_h)\}$$

and

$$V(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1}\left[E\left\{\hat{V}^*_\infty(\hat{Y}_h)\right\} - \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K\right].$$

Replacing $E\{\hat{B}^*_\infty(\hat{Y}_h)\}$ and $E\{\hat{V}^*_\infty(\hat{Y}_h)\}$ in the right-hand sides by their respective (unbiased) estimators $\hat{B}^*_R(\hat{Y}_h)$ and $\hat{V}^*_R(\hat{Y}_h)$, we obtain Expressions (7) and (8). We can also obtain analytical versions of these bias-corrected bootstrap estimators by using (13) and (14):

$$\hat{B}^*_{\infty,BC}(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1}(1-p)\left\{\overline{\hat{Y}}^{(-h)} - \hat{Y}_h\right\}, \tag{17}$$

$$\begin{aligned}\hat{V}^*_{\infty,BC}(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1}\Bigg[&p(1-p)\hat{K}_h + \frac{1-p}{H-1}\left(1 - \frac{1-p}{H-1}\right)\sum_{g\neq h}\hat{K}_g \\ &- \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K\Bigg].\end{aligned} \tag{18}$$

6. References

Bakker, B.F.M. and P.J.H. Daas. 2012. "Methodological Challenges of Register-Based Research." *Statistica Neerlandica* 66: 2–7. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00505.x.

Bethlehem, J. 2009. *Applied Survey Methods: A Statistical Perspective*. Hoboken, NJ: Wiley.

Bryant, J. and P. Graham. 2013. "A Bayesian Method for Deriving Population Statistics from Multiple Imperfect Data Sources." Paper presented at the World Statistics Congress, August 25–30, Hong Kong. Available at: http://www.statistics.gov.hk/wsc/IPS027-P4-S.pdf (accessed December 2013).

Burger, J., J. Davies, D. Lewis, A. Van Delden, P. Daas, and J.-M. Frost. 2013. *Guidance on the Accuracy of Mixed-Source Statistics. Deliverable 6.3/2011 of ESSnet Admin Data*. Available at: http://essnet.admindata.eu/WorkPackage/ShowAllDocuments?objectId=4257 (accessed December 2013).

Chamberlain, J. and E. Schulte Nordholt. 2004. "The Results of the 2001 Census in the Netherlands, the United Kingdom and Some Other European Countries." In *The Dutch Virtual Census of 2001, Analysis and Methodology*, edited by E. Schulte Nordholt, M. Hartgers and R. Gircour, 225–241. Statistics Netherlands: Voorburg/Heerlen.

Delden, A. van and P.P. de Wolf. 2013. "A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data." In Proceedings of the Conferences on New Techniques and Technologies for Statistics, March 5–7 2013. Brussels. Available at: http://www.cros-portal.eu/sites/default/files//NTTS2013%20Proceedings_0.pdf (accessed December 2013).

Demnati, A. and J.N.K. Rao. 2009. "Linearization Variance Estimation and Allocation for Two-Phase Sampling under Mass Imputation." Paper for the Federal Committee on Statistical Methodology Research Conference, November 2–4, Washington, DC. Available at: http://www.fcsm.gov/09papers/Demnati_VI-C.pdf (accessed December 2013).

De Munnik, D., M. Illing, and D. Dupuis. 2013. "Assessing the Accuracy of Non-Random Business Conditions Surveys: a Novel Approach." *Journal of the Royal Statistical Society, Series A*, 176: 371–388. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01035.x.

DiCiccio, T.J. and B. Efron. 1996. "Bootstrap Confidence Intervals." *Statistical Science* 11: 189–228. Doi: http://dx.doi.org/10.1214/ss/1032280214.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

Kuijvenhoven, L., and S. Scholtus. 2011. *Bootstrapping Combined Estimator Based on Register and Sample Survey Data*. Discussion paper 201123. The Hague/Heerlen: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/06202B2A-B6C1-40CC-B25B-4022B7712E59/0/2011x1023.pdf (accessed December 2013).

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

Shao, J. and R.R. Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association* 91: 1278–1288. Doi: http://dx.doi.org/10.1080/01621459.1996.10476997.

UNECE. 2007. *Register-Based Statistics in the Nordic Countries, Review of Best Practices with Focus on Population and Social Statistics*. New York: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed August 2015).

Zhang, L.-C. 2011. "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics* 27: 415–432.

Zhang, L.-C. 2012a. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Zhang, L.-C. 2012b. "On the Accuracy of Register-Based Census Employment Statistics." Paper presented at the European Conference on Quality in Official Statistics (Q2012), May 30–June 1, Athens. Available at: http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf (accessed December 2013).

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 507–513, http://dx.doi.org/10.1515/JOS-2015-0030

Discussion

Ray Chambers[1]

1. Introduction

I am very grateful for the opportunity to contribute to this special issue of the Journal of Official Statistics by commenting on the articles in it. In particular, I have chosen to focus my comments on the articles by Burger et al., Gerritse et al., Tuoto and Di Consiglio, and Zhang, because these authors, to a greater or lesser extent, tackle measurement-error issues that are important emerging features of official statistics methodology.

2. Comments on Burger et al. article

I start with the article by Burger et al. This addresses the important issue of industry misclassification when records from a survey and an administrative data source are combined. In particular, the article considers a business survey application where in fact a census is carried out, in the sense that there a 100% survey of large businesses is conducted, with data for the remaining medium and small businesses extracted from a tax register. To quote the authors, "Because no samples are drawn and missing data are imputed, no complicated design-based or model-based estimators are required to make inference about the target population." This of course ignores the whole minefield of imputation bias and variability, as well as the usual conceptual issues that arise when two variables ostensibly referring to the same thing are measured in two different ways. But once one pushes this (huge) elephant out of the living room, then the issue of errors in the industry classification of the units in the two sources can be considered. The article introduces a simple model for misclassification errors within a group of industries that is the same as the simple exchangeable model for linkage errors introduced by Neter et al. (1965), and used as the basis for bias correction in that context in a series of papers starting with Chambers (2009). However, the authors of this article are not interested in bias correction *per se*, focusing instead on bootstrap simulation of the extent of the bias and the increase in variability that arise under a multinomial version of this simple model. Here their results are sobering, indicating quite significant increases in both bias and variability even when the data meet the quality specifications of an internal Service Level Agreement (SLA) on classification accuracy. Interestingly, the results in the article show that because higher levels of accuracy in classifying small to medium businesses lead to reductions in bias relative to expected levels under the SLA, there is in fact a large bias-variance trade-off to be made in terms of allocating resources for carrying out the classification. No information is provided on how this trade-off can be (was?) eventually resolved, but, again quoting the authors, "the current paper provides insight into the sensitivity of mixed source statistics to a source-specific nonsampling error." Much more research needs to be done,

[1] National Institute for Applied Statistics Research, Wollongong NSW 2522, Australia. Email: ray@uow.edu.au

particularly in terms of developing robust misclassification bias corrections for the outputs from the application. This is particularly the case since these outputs appear to be an important component of the information used in determining gross domestic product. In this context, the work on bias correction for linkage errors may prove useful, see Kim and Chambers (2012).

3. Comments on Gerritse et al. article

The remaining three articles all focus on a different type of measurement error, introduced when two or more data sources, each with incomplete coverage of a population of interest, are linked in order to estimate the total population size. This of course is the classical census coverage problem, and the so-called dual-system estimation (DSE) methodology for dealing with it is now well established. The article by Gerritse et al. uses the DSE as a jumping-off point, providing a nice overview of the main issues that arise when using this approach, and particularly focusing on the problems that arise when the union of the two sources is a subset of the population of interest (so undercoverage is the focus) and the key assumption of independent coverage errors for the two data sources is in fact incorrect. In this context it helps to introduce some notation, so let $A = 1(0)$ denote the event that a population unit (of some agreed type) is included (not included) in the first data source, and let $B = 1(0)$ denote the same two events for the second data source. Put $N_A(N_B)$ equal to the known counts of population units with $A = 1 \ (B = 1)$ and put L equal to the set of linked units, with X_{11} equal to the linked count, that is the number of units with $A = 1$ and $B = 1$. The DSE for the unknown total population size N is then $(N_A \times N_B)/X_{11}$, and can be easily shown to be the method of moments estimator for N under a number of assumptions, a crucial one of which is independent 'capture' events for the same population unit relative to the two data sources.

There have been a variety of suggestions in the literature on reducing the bias that ensues when the two data sources are in fact not independent. However, as the authors emphasise, "Independence is an unverifiable assumption, that is, it cannot be verified from the data used for the estimation of the population size." Consequently, given the available data, all one can do is carry out numerical exercises based on the data at hand to demonstrate sensitivity to failure of this assumption, or carry out studies to investigate bias under simulated conditions. Following Brown et al. (1999, 2006) these authors take the first approach and investigate the sensitivity of the DSE estimates obtained by linking records on the Dutch Population Register with records on a police register. Like Brown et al. (2006), the approach is based on perturbing the odds ratio in a log-linear model for the complete cross classification of the target population, though the methodology presented in the article extends this model to one of Poisson counts and also considers the case where heterogeneous coverage probabilities arise because of covariate information from one or both of the data sources. As one would expect, the higher the achieved coverage, the less sensitive are the DSE-based methods to break down in the independence assumption. This is nicely illustrated in the application described in the article, where a realistic variation in the odds ratio leads to biases in the range -15% to $+9\%$ for the estimated counts of people of either gender and with an Afghan, Iraqi, and Iranian nationality two years previously, compared with biases in the range -42% to $+58\%$ for the corresponding estimates of people with Polish nationality. As the authors point out, the main reason for this difference is the fact that Dutch and EU law ensure that the overall coverage of the first group by the two data sources is much higher than the corresponding coverage of the second group. However, the fact that such biases can occur is a salutary reminder that failure of model assumptions can have a much more dramatic impact when one is dealing with measurement error than, for example, when one is using regression models for prediction in a 'pure' sample-survey context.

4. Comments on Tuoto and Di Consiglio article

Turning now to the article by Tuoto and Di Consiglio, we see that these authors consider exactly the same situation as that considered by Gerritse et al. but in this case focus on a different measurement-error problem, that of linkage errors when the two data sources are integrated to obtain X_{11}. These authors also use a different nomenclature from that used in Gerritse et al., referring to the DSE estimator as the Petersen estimator, reflecting its origin in estimating the sizes of wild animal populations in the late nineteenth century. As in Gerritse et al., there is an (unspoken) assumption of multinomial sampling throughout, allowing the straightforward development of estimators from moments of unknown quantities. In addition to the definition of L and X_{11}, define $A - L$ as the set of X_{10} population units on A but not on B, that is, X_{10} is the number of records found to be only on list A. Similarly, define $B - L(X_{01})$ to be the set (number) of records found to be only on list B. Then $N_A = X_{11} + X_{10}$ and $N_B = X_{11} + X_{01}$. Under independence and perfect linkage,

$$E(X_{11}) = N\,\Pr(\text{record in } A)\,\Pr(\text{record in } B)$$

while

$$E(N_A) = N\,\Pr(\text{record in } A) = N\tau_1$$

$$E(N_B) = N\,\Pr(\text{record in } B) = N\tau_2$$

so, using a 'hat' to denote an estimate,

$$\widehat{\Pr}(\text{record in } A) = \hat{\tau}_1 = X_{11}/N_B$$

$$\widehat{\Pr}(\text{record in } B) = \hat{\tau}_2 = X_{11}/N_A$$

and therefore, setting $M = X_{11} + X_{10} + X_{01}$, we have $E(M) = N(\tau_1 + \tau_2 - \tau_1\tau_2)$. The Petersen estimator of N follows by replacing the unknown parameters in this expression by their moment estimates, leading to

$$\hat{N} = M/(\hat{\tau}_1 + \hat{\tau}_2 - \hat{\tau}_1\hat{\tau}_2)$$

It is straightforward to see that this estimator is identical to the DSE defined earlier.

However, the reality in most cases is that there are errors in linking, in the sense that records common to both lists are not matched, as well as matched records that are incorrectly matched. This problem is (partially) addressed by Ding and Fienberg (1994), who assume incorrect matching is only from A to B. In this context, one can define

$$\alpha = \Pr(\text{correct match}) = \Pr(\text{match is a record from } L)$$

$$\beta = \Pr(\text{incorrect link}|\text{match}) = \Pr(A - L \text{ record matched to } B \text{ record})$$

It follows that

$$\Pr(L \text{ unit linked}) = \alpha \Pr(L \text{ unit}) = \alpha\tau_1\tau_2$$

$$\Pr(A - L \text{ unit linked}) = \beta \Pr(A - L \text{ unit}) = \beta\tau_1(1 - \tau_2)$$

$$\Pr(B - L \text{ unit linked}) = 0$$

and so

$$E(X_{11}) = N(\alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2))$$

$$E(X_{10}) = E(N_A) - E(X_{11}) = N(\tau_1 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2))$$

$$E(X_{01}) = E(N_B) - E(X_{11}) = N(\tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2))$$

Since a population unit that is not on either data set cannot be matched to one that is, it follows that $M = X_{11} + X_{10} + X_{01}$ is the number of unique population units identified in the union of the two data sources, with

$$E(M) = N(\tau_1 + \tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2))$$

Assuming estimates of α and β are available from the linking process, the Ding and Fienberg estimator of N is the method of moments estimator derived from this identity, with τ_1 and τ_2 replaced by their moment-based estimates, which must then satisfy

$$\hat{\tau}_1 = N_A/\hat{N} = (N_A/M)(\hat{\tau}_1 + \hat{\tau}_2 - (\alpha - \beta)\hat{\tau}_1\hat{\tau}_2 - \beta\hat{\tau}_1)$$

$$\hat{\tau}_2 = N_B/\hat{N} = (N_B/M)(\hat{\tau}_1 + \hat{\tau}_2 - (\alpha - \beta)\hat{\tau}_1\hat{\tau}_2 - \beta\hat{\tau}_1)$$

Solving for $\hat{\tau}_1$ and $\hat{\tau}_2$ based on these identities, we obtain

$$\hat{\tau}_1 = (X_{11} - N_A\beta)/(N_B(\alpha - \beta))$$

and

$$\hat{\tau}_2 = (X_{11} - N_A\beta)/(N_A(\alpha - \beta))$$

It is straightforward to see that in the case of no linkage error, that is $\alpha = 1$ and $\beta = 0$, the Ding and Feinberg estimator defined by $E(M)$ above reduces to the Petersen estimator.

The article by Tuoto and Di Consiglio extends this idea to also allow linkage errors from B to A. In order to do this, these authors assume that the probability of this happening is the same as the probability of incorrect matching from A to B (i.e., β). Then, following the same approach as that underpinning the Ding and Fienberg estimator, it can be seen that

$$E(X_{11}) = N(\alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2) + \beta\tau_2(1 - \tau_1))$$

$$E(X_{10}) = E(N_A) - E(X_{11}) = N(\tau_1 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2) - \beta\tau_2(1 - \tau_1))$$

$$E(X_{01}) = E(N_B) - E(X_{11}) = N(\tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2) - \beta\tau_2(1 - \tau_1)),$$

so collecting terms

$$E(M) = N(\tau_1 + \tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2) - \beta\tau_2(1 - \tau_1))$$

The same argument as used by Ding and Fienberg then leads to

$$\hat{\tau}_1 = N_A/\hat{N} = (\beta M + X_{11}(\beta - 1))/(N_A(2\beta - \alpha))$$

$$\hat{\tau}_2 = N_B/\hat{N} = (\beta M + X_{11}(\beta - 1))/(N_B(2\beta - \alpha)).$$

Substitution of these expressions for $\hat{\tau}_1$ and $\hat{\tau}_2$ into the method of moments estimator of N defined by the preceding expression for $E(M)$ leads to the adjusted estimator for N defined by Expression (13) in the article.

As noted by Tuoto and Di Consiglio, the main advantage of (13) over the standard Ding and Fienberg approach is bias reduction when β is non-negligible. However, this assumes symmetry of incorrect matching between A and B, which is debatable and should be possible to generalise. Also, the approach depends on having access to good estimates of linkage-error probabilities, which can require audit samples. In this context it is important to note that these values of α and β must be such that the estimate $\hat{N}$ of N defined by (13) in the article satisfies the consistency restrictions defined by the Fréchet inequalities,

$$\max(N_A, N_B) \leq \hat{N} \leq \min(N_A, N_B)(\alpha\hat{\tau}_1\hat{\tau}_2 + \beta\hat{\tau}_1(1 - \hat{\tau}_2) + \beta\hat{\tau}_2(1 - \hat{\tau}_1))^{-1}.$$

5. Comments on Zhang article

Finally, I turn to the article by Zhang. This considers another possible source of measurement error when a population size is estimated by linking two or more data sources. In this case the author tackles the situation where two population lists (or registers) are linked in order to estimate the size of a population that is partially captured by each list. The twist is that these lists also include units that are not from the population of interest. In other words, there is both undercoverage as well as overcoverage when the two lists are linked. We can characterise this situation using the schematic below. This shows a target population U of (unknown) size N, partially covered by two linked lists, denoted as usual by A and B. Without loss of generality we denote membership of $A(B)$ by

	$U = 1$		
	$B = 1$	$B = 0$	
$A = 1$	N_{11}	N_{10}	N_A
$A = 0$	N_{01}	N_{00}	$N - N_A$
	N_B	$N - N_B$	N
	$U = 0$		
	$B = 1$	$B = 0$	
$A = 1$	K_{11}	K_{10}	K_A
$A = 0$	K_{01}	0	$K - K_A$
	K_B	$K - K_B$	K

the binary event $A = 1$ $(B = 1)$. Similarly, membership of U is denoted by the binary event $U = 1$.

The author refers to the set of $N + K$ units covered by this schematic as the target-list universe and assumes an underlying multinomial distribution for the cell counts defining it. Note the structural zero for the (000) cell, since the target-list universe cannot contain such units. The author also assumes

- An independent coverage survey with only undercoverage error (all surveyed units are $U = 1$) but with unknown target population coverage. That is,

  $$\pi = \Pr\left(\text{unit in } U \text{ included in sample}\right)$$

  is unknown. This will be the case if the framework used to select the sample for the coverage survey is a subset of U.

- Perfect linking of A and B as well as linking of coverage survey units to A and B. Consequently, $X_{11} = N_{11} + K_{11}, X_{01} = N_{01} + K_{01}$ and $X_{10} = N_{10} + K_{10}$ are known, as is the corresponding breakdown of the survey counts, which we denote n_{11}, n_{10}, n_{01} and n_{00}, with the usual interpretation.

Note that there is no assumption of independence between A and B. The aim is to use these data to estimate N.

Let τ_{jk} denote the conditional probability that a randomly sampled unit from the target-list universe has $A = j$ and $B = k$ given that it is a member of the target population, that is, has $U = 1$. Then, under the assumed multinomial model for the target-list universe, the linked list counts satisfy $E(N_{jk}) = X_{jk}\tau_{jk}$, and for the corresponding linked sample counts, $E(n_{jk}) = X_{jk}\tau_{jk}\pi$, with

$$E(n_{00}) = E(N)\pi - E(n_{11}) - E(n_{10}) - E(n_{01})$$

Unfortunately, without knowing the value of π, the equation for $E(n_{00})$ above shows that the available data are insufficient to identify N given the assumed multinomial model for the target-list universe. Another identifying assumption is needed. In the article, the author uses a log-linear model characterisation of the problem to investigate alternative approaches to resolving this identification problem, with the most promising of these based on a ‘pseudo-independence’ assumption for the list universe defined by the union of A and B. This is where the probability of a nontarget population unit in this universe being linked is the product of the corresponding probabilities of a nontarget population unit being on either list, see Equation (11) in the article. The author argues that this assumption is reasonable when the lists are of high quality, that is, there are few target population units missed by them, and derives the method of moments estimators of these probabilities, see Equation (13). The corresponding method of moments estimator of N then follows from standard arguments.

6. Some Concluding Observations

From the perspective of a commentator, all four articles reviewed above have a common focus. They all consider problems that arise when situations corresponding to nonstandard measurement error scenarios arise in official statistics. The way they tackle these problems is different. The first two articles, by Burger et al. and Gerritse et al., use sensitivity analysis and simulation to illustrate the extent of the problem when standard statistical methods (which ignore the measurement error) are used. As we see, their findings are sobering. The glass is definitely half empty. The articles by Tuoto and Di Consiglio and by Zhang are more along the lines of the glass being half full. Both focus on remedial action, extending the models underpinning the standard methods to accommodate the measurement error. Their results are encouraging, in the sense that they show that these errors can be dealt with in a systematic way. However, they are far from being the final word on the matter. Both tackle the estimation problem, but leave the (hard!) inference problem for later. The reason for this is clear – unlike the well-known sample error structure that is implicit in conventional official statistics, modern official statistics is increasingly eschewing sampling or minimising the use of (expensive) samples, instead using a variety of linking and combining techniques to create what is hopefully something like a 'census' of the population of interest. As these authors clearly demonstrate, this can be a fool's paradise. The errors implicit in linking (or even more importantly, nonlinking), as well as misspecification errors in the implicit models underpinning the estimates derived from these data, can be considerable. The four articles in this issue that I have commented on here represent significant steps towards development of a methodological framework for inference in such situations. It is quite obvious that such a framework will depend on modelling assumptions, so the classical design-based inference paradigm that has for so long served so well in official statistics is irrelevant. What we see here is evidence that the model-based inference paradigm for official statistics that is taking its place needs to be applied with a strong dose of common sense, and a good knowledge of the frailties of the models used. The insurance provided by design-controlled randomisation is no longer available.

7. References

Brown, J.J., O. Abbott, and I.D. Diamond. 2006. "Dependence in the 2001 One-Number Census Project." *Journal of the Royal Statistical Society Series (Statistics in Society)* 169: 883–902. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00431.x

Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. "A Methodological Strategy for a One-Number Census in the UK." *Journal of the Royal Statistical Society Series A* 162: 247–267.

Chambers, R. 2009. "Regression Analysis of Probability-Linked Data." *Official Statistics Research Series*. Available at: http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

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

Kim, G., and R. Chambers. 2012. "Regression Analysis Under Incomplete Linkage." *Computational Statistics and Data Analysis* 56: 2756–2770.

Neter, J., E.S. Maynes, and R. Ramanathan. 1965. "The Effect of Mismatching on the Measurement of Response Error." *Journal of the American Statistical Association* 60: 1005–1027.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 515–525, http://dx.doi.org/10.1515/JOS-2015-0031

Discussion

Anders Holmberg[1]

1. Introduction

I would like to thank the editors for the opportunity to comment on the coverage issues affecting administrative data (AD) in this special issue of *The Journal of Official Statistics*. I will follow the definition provided in UNECE (2011) and refer to AD as data collected external to statistical offices, while administrative sources are data holdings that contain information not primarily for statistical purposes, either private or public. My definition of the noun 'survey' includes research that is designed and based on statistics from such sources. Hence, an AD survey or integration survey lacks purpose-built questionnaires, and its original data-acquisition instruments are outside the full control of statistical offices and researchers.

Methodology research for statistics mainly using AD has picked up pace and this special issue demonstrates this fact. One reason is increased worldwide interest in using AD in population censuses. In the last European census, some countries moved away from a traditional census. Others, such as the United Kingdom, New Zealand, the United States and Canada have ongoing census modernisation programs containing significant efforts to investigate the use of AD. However, this interest is not completely new. Scheuren (1999) and the references therein illustrate that it was on the agenda in the US as far back as in the 1980s. Another reason may be that the geographical spread and collaboration between National Statistical Organizations (NSOs) and academia have created a critical mass. Not too long ago, methodological work on AD were restricted to fragments inside NSOs, and in the field of social statistics it was practiced mainly by the Nordic countries, the Netherlands and Slovenia (e.g., see Nordbotten 1966; UNECE 2007; Schulte-Nordholt et al. 2004; Zaletel and Krizman 2008). It is therefore pleasant to see the mix of countries represented in this issue.

If I ignore AD used as auxiliary information in the design and estimation of sample surveys, my personal experience with AD goes back approximately twelve years. During this time I worked with AD methods in business statistics, social statistics, and a register-based census, as well as trying to facilitate an organisational view to improve the use of registers and AD in a national statistics production system. I will reflect on this period and provide some ideas about AD and statistics that have become 'food for thought' after reading these articles, and which (in my opinion) need attention.

My discussion will not focus on the articles' details, but instead make a note of their fit with NSO activities, bearing in mind that NSOs today not only make statistics, some also provide microdata for researchers as a part of their countries' data infrastructure. These infrastructures, which consist largely of AD, are significant and contribute to unlocking the

[1] Statistics New Zealand, PO Box 2922, Wellington 6140, New Zealand. Email: anders.holmberg@stats.govt.nz

value of data – in a safe and trusted way. NSOs have a great opportunity to combine infrastructures for microdata with modernised statistics systems. As public and private AD sources grow, it is vital to align the production systems of official statistics with these infrastructures, with new statistics applications, and with the development of statistical methods. This JOS issue deals with some methodological challenges that follow, namely coverage, linking methods, and subsequent estimation. The estimation techniques proposed will mean that statistical modelling and computer-intensive methods must increase in use. I intend to discuss some points about the opportunity (and challenge) facing NSOs based upon my experience of AD from statistical offices in Sweden, Bolivia, Cambodia, and New Zealand.

2. Data Integration and AD Surveys

As ‘flagships’ such as population censuses radically change design, it is becoming clearer that the field of survey design is gradually shifting. It is moving from (albeit complex) sample surveys to surveys based on integrated data with AD as a backbone. NSOs that realise and adapt to this change face both an opportunity and a challenge. The opportunity is to use their responsibility and participation to build national data infrastructures and create production environments enhancing integrated survey statistics. As foreseen by Nordbotten and Scheuren, this is cost efficient from a societal perspective and complements sample survey programs by delivering broader, more detailed, and more responsive subject-matter contents. A production environment for integrated data and multiple source statistics would also enable NSOs to play an active role shaping new and alternative data sources and collection methods. Their main challenge is to align the workforce and the production processes.

For NSOs, this means that an end-to-end statistics production process will rely more on data streams with different origins. Production environments must also be able to effectively and efficiently exploit the possibilities of data integration. When doing this, it is necessary to have secure and well-designed IT systems for storage, processing and access, but sound methods are even more essential. NSOs that try to modernise their end-to-end processes with little or no thought to survey designs for data integration risk making bad investments in inadequate IT structures.

2.1. Statistical Modelling and Validation Efforts Will Increase

My first encounters with statistics that relied solely on AD were in business statistics through projects on improving timeliness and accuracy. These projects had only one main data source, which had only one specific use. The tasks therefore resembled those of improving a single-purpose sample survey. Despite the main goal of improving timeliness, the projects spent little effort on data acquisition processes. Instead, the focus was on developing estimation techniques that could provide *rapid (preliminary) estimates* that were robust against bias caused by measurement errors and missing units. Just as in this special issue, statistical modelling played a crucial part.

Six of the articles in this special issue present estimation techniques based on statistical modelling. Five of them (Zhang, Gerritse et al., Chipperfield and Chambers (C&C), Yildiz and Smith (Y&S) and Di Consiglio and Tuoto (D&T)) discuss log-linear models, and one article, Bryant and Graham (B&G), discusses Bayesian techniques. This is not surprising and it is safe to predict that if more AD is used, all forms of statistical models will play a greater role in official statistics. For NSOs, the challenge would be to explain to users the necessity of models and their impact on statistics quality, particularly when there are many model variants to choose from and different statisticians to trust. It is important to pursue ways of validating model assumptions and estimating errors caused by their violation. I refer to Gerritse et al. as a valuable contribution in this respect. Because of higher recurrence, my experience is that it is easier to validate models in economic statistics than in social applications. One of the abovementioned projects was carried out on monthly statistics, and before introducing a new method we monitored patterns of incoming data over several rounds. We were able to repeatedly compare preliminary estimates based on incomplete data with corresponding final estimates and thereby empirically check competing estimation models (Jäder and Holmberg 2005). This method is not practical with less frequent data collections and definitely not for models proposed for a census. In this case, other validation methods are necessary that may incorporate extra data collections and/or experiments and add to cost. Y&S and B&G give two very different census estimation methods using models. NSOs considering these should look for ways to compare them, which is not straightforward.

2.2. Linking and Microdata Access

It is typical for modern statistics using AD to reuse data through integrating and combining different sources. I first came across multiple uses and integration when I worked with Statistics Sweden's Microdata ONline Access system for researchers (MONA). This system contains primarily personal data and has a design that is far more *ad hoc* than the data archives solution advocated by Nordbotten (1966). In MONA, personal identification numbers are available and they provide unique unit record identifiers, which make data integration and high-quality record linkage easy.

Internationally this is unusual – in many environments record linking is a major undertaking that requires significant methodological effort. C&C, D&T and Blackwell et al. illustrate this with different linking aspects. The first two authors present estimation methods in the presence of imperfect linking. Blackwell et al. illustrate the complexities and practical barriers that exist in a big project, such as linking census data with AD. Because of varying circumstances, it is probably unwise to copy Blackwell et al.'s approach exactly. However, the article shows a range of necessary steps and available possibilities by mixing exact/deterministic matching with probabilistic and clerical routines. All this is done to maximise linking rates with as few errors as possible.

Describing the size of the linkage error and compensating for it is indeed a methodological task. Estimates of the true positive rates (the sensitivity) and true negative rates (the specificity) should routinely accompany any linked data. Still, the set of negative links rarely gets the attention it deserves. It is worth looking closely at the records that do not link. This should give good insights into AD patterns, as the false negatives (whenever detected in reviews) are similar to studying the attributes of nonrespondents in a sample survey. The true negatives may reveal other deficiencies in the AD sources – coverage is one of them.

C&C, the references therein, and to some extent D&T, present methods for handling the effect of a certain type of linkage error. Demand for using these methods will increase as a result of NSOs creating research analysis infrastructures with linked microdata. Statistics New Zealand's Integrated Data Infrastructure (IDI) is one of these interesting environments under development. It allows for statistical outputs and research on the transitions and outcomes of *people* through various areas. With a conscious approach to confidentiality and security, the IDI provides analysts with microdata that sometimes are the result of linking multiple datasets. On top of the abovementioned quality traits for linking, transitivity is then introduced as another concern. Blackwell et al. have only one AD source, but NSOs that might, for coverage reasons, want to combine multiple AD sources before linking should study transitivity effects (Sadinle and Fienberg 2013).

2.3. Coverage and Statistical Units in Production Environments for Integrated Data

The raw records of many AD sources in MONA and IDI are based on registered events, or (if there is no terminating event) a relation between entities, for example employer/employee, hospital/patient, school/pupil. The records are usually transformed into units of interest such as persons, but sometimes, depending on the purpose, they are kept in their original form as records of employment, treatment, course enrolment and so on. Zhang (2012) uses base units and composite units as a way of understanding the quality properties of integrated AD. This is a useful distinction in studying the interplay between coverage issues and linking, since coverage is defined by the target unit and that unit is not necessarily the linking unit. Linkage errors have a direct effect on coverage, whether the linking unit is the target unit or not.

At the integration/linking stage, reasons other than linkage errors can influence coverage. Zhang's model introduces an alignment stage to sort the relations between base units and composite units in integrated data. It also introduces identification errors and unit errors that are conceptually different but where the effects are similar to those of coverage errors. Burger et al. treat this when they study the effects of setting a single industry code for a composite unit, such as an enterprise unit, when its LKAUs (Local Kind of Activity Unit) have different industry codes. The Swedish register-based census is another recent example where coverage problems arise because of unit errors in integrated data. The post-census evaluation survey indicates that the register-based sources for the census underestimate the number of one- and two-person households and overestimate the number of households with six or more members. Since person coverage is good, the overall effect is an underestimation of the total number of households by 4-5 percent depending on domain (see Andersson et al. 2013). Hence, good coverage of the base unit (person) does not mean good coverage of the composite (household) unit. With access to a greater variety of data sources containing different unit types, NSOs need good functions to handle coverage errors and other problems arising from the integration stage.

I think a flexible and cohesive system for data integration is easier to achieve if the statistical business architecture is built around appropriate base and composite units. Most statistics about society have units related to land, people, or business. In these three spheres, AD is usually available from the public sector. Hence, with legal access or even custodianship of such core AD, the NSOs have better opportunities than others to sort appropriate statistical units, to standardise the units and to build good infrastructures for multiple-source statistics with such units as a backbone.

Figure 1 shows a simple unit-centric structure with relations between important statistical units in the subject spheres of land, people and business. Complemented with methods for treating the units' time and geographical dimensions, it is a foundation for defining and accessing target units and for applying data-integration methods from a statistical system's perspective. The keystone units in the illustration are base as well as composite units, and in the case of dwellings are both, depending on the statistics question at hand. Dwelling unit is included here to show a unit that establishes a connection between people and land through household/housing statistics. Otherwise each sphere can be expanded and has a set of units not shown here for simplicity. (For example, in a detailed picture the business sphere would have Kind of Activity Units (KAU), local KAUs and legal units – and, if it helps, enterprise groups. The land sphere would have building and entrance units and the people sphere would have household and family units.) In a system structure for integrated data, the geographical attributes in the middle are very important. They are central to the integration apparatus (especially without well-established identifiers) and should not be used only for statistical collection and dissemination processes. Also, by expanding the unit-centric structure below it is also fairly straightforward to put context to and interpret event/activity records as relations between units. A lot of useful AD statistics are based on such data.

Storage, access, and maintenance of the unit data can be done in statistical registers, as described by Wallgren and Wallgren (2007). This can also be done in other ways, for example a system of unit frames which are tied together by a linking methodology and effective data processing capabilities. The unit-centric approach facilitates the development of an environment that can integrate data quickly in a standard, transparent, and interpretable way. A huge benefit is that it enables assessments of various target and accessible survey populations. It also simplifies the interlinking of different subject-matter areas and makes it easier to assess which data are best in a multisource choice situation. The populations in turn can be national benchmarks with well-known coverage properties

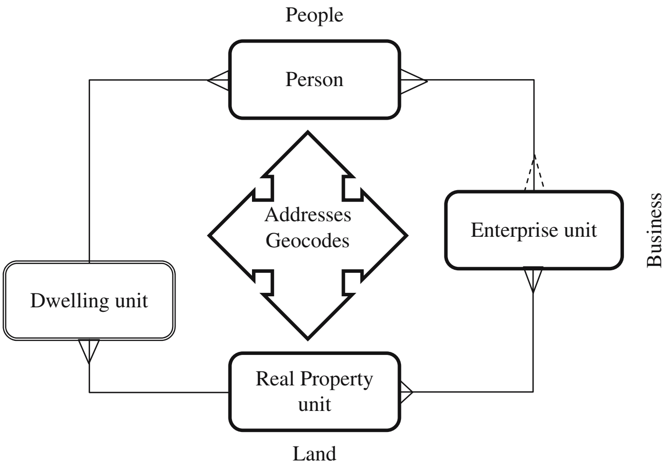

Fig. 1. A unit-centric statistical structure for integrated data surveys

to be used by many in comparative studies as well as official statistics. They can also give meaning to pointless statements against the sampling paradigm such as "...gathering as much as possible, and if feasible, getting everything: N = all" found in the big data literature (Mayer-Schönberger and Cukier 2013, 29). Ultimately, without unit understanding and a sought population it is hard to evaluate what is meant by "all"; surely there are cases when you get more than all. With methodological know-how, NSOs can make sense of integrated data by putting them in context, explaining coverage after linking, and perhaps also improving the quality of AD systems.

3. Development Areas for Integrated Surveys and AD Systems

In this section I highlight some other development areas I considered while reading this special issue and while thinking about how NSOs work with AD.

3.1. Expand the Methods Toolbox Using Geographical AD

Developing a structure such as Figure 1 means we must pay more attention to geographical AD and the location concept. While many NSOs are good at conforming to the geodata evolution when they disseminate statistics, it is still more or less uncharted territory for methodologists designing surveys or working generally with AD. There has been progress in the traditional use of AD, such as standardised solutions to communicate with GIS systems and map the hierarchies of areas relating to national and local administrative geographies, but NSOs seem slow to take up new statistical methods with geospatial data.

I suspect that soon we will see more integration surveys based on geographical linking. These surveys will be based not only on addresses (which require substantial cleaning efforts) but also on geocodes, clusters of geocodes using geohashes and 'snap-to-grid' methods (Heath and Goodwin 2011). Naturally, geographical linking requires good geocoding practices when the AD are created. This already exists partly, both in public and private data, but NSOs should be ready to take advantage of this and regularly add geocodes to their own data collections. This enables easier and more reliable linking between units of different types. It also allows the creation of new types of geodependent composite units.

To give an example, in Sweden practically all electricity meters are geocoded for reasons of repair and reading. The meters are connected to dwellings rather than buildings. By using a geocode link (as one example among others), the chance to infer dwelling occupancy based on electricity consumption is good. This is an interesting option in population censuses and housing statistics. With slight adaption, the ideas in this special issue should be applicable for errors using geographical linkage; the linking articles C&C and D&T are particularly interesting, as is Burger et al.'s contribution. As far as I understand, the classification/coverage problem they treat can also be adapted to composite 'proxy' units linked together by geography. Sometimes you want to classify aggregated composite units (e.g., geographically linked groups of buildings, dwellings or households) that have diverging information on the base-unit level. In another setting, Burger et al.'s approach may also clarify the sensitivity of classification errors on association measures applied on geodependent composite units. Linking composite units is also possible when a base unit option is hard to get or not allowed because of legal constraints.

3.2. Examine Time Dimension in AD Systems and Analyse Events and Delayed Data

Time dimension is a critical factor for the coverage and linking of AD. In some sources it can be tricky to distinguish between reference dates and registration dates. There also must be operational solutions for how to relate the data to time, for example the usual residence at a single point in time. Since many AD sources have records that are events or relations and since storage and processing systems often are poorly designed regarding the *statistical* units, studies on units' status-change frequency are rare. A lot can be learned about an AD source by consciously monitoring and analysing unit changes. Changes are not only signals of underlying societal and population changes; they can also be signs of alterations of administrative routines in the source. Moreover, provided that historic or change data are kept, some of the AD retained by NSOs have longitudinal information waiting to be unearthed by computer-intensive pattern recognition methods.

Event or delayed data are also potential sources that can help us to understand how coverage evolves over time. It is not unusual for delayed data phenomena to appear in recurrent business surveys as a survey feedback issue. Often the recommendation is to ignore the information since it introduces estimation bias. However, this practice also neglects coverage errors and the trade-offs are not always straightforward. Delays can sometimes also prevent accurate linking.

Other NSO activities can also benefit from event data in AD. Every day the Swedish population register gets updates on events such as address changes, changes in marriage status, births and deaths. If changes (e.g., moving house, divorce) make people harder to reach, it makes sense to transfer or at least compare this information with that from surveys doing collection and estimation. With survey designs using direct element sampling and a mixed-mode mail and web or CATI collection, this may reduce a significant part of the nonresponse due to no established contact, or reduce bias in calibration estimation.

3.3. Measuring Coverage, Coverage Targets and Estimation

This special issue and the work in the Beyond 2011 program (run by the Office of National Statistics to investigate alternative census possibilities in England and Wales, see ONS 2013; Skinner et al. 2013) reveal a focus shift in viewing census coverage. Undercoverage is the most serious issue in a traditional census, and post-census surveys are designed to deal with this, usually through area sample designs that are independent of census collection. However, in a census based on AD, both overcoverage and undercoverage seem likely. These are not expected to be evenly distributed. On the contrary, just cross-examining AD over geography, sex and age is likely to produce complicated patterns of included and excluded units. Therefore it might prove difficult to estimate the extent of both types of coverage errors efficiently using one single survey. The underlying AD mechanisms of the coverage problems can be very different, which is well illustrated by the data in Gerritse et al. In that context it makes sense to view post-census activities as a package of actions with maybe more than one data collection. The practices around post-census data collections and their implications on estimation methods need to be updated, and the solutions are connected to the choice of a dual-system estimation method or perhaps even triple-system estimation as discussed by Griffin (2014). The independence assumption between sources is highlighted by Zhang and Gerritse et al., and is also considered by Y&S and B&G.

The coverage issue in AD arises when sources are used for statistical purposes. It is the obvious cause of error to study when considering AD because its effect is easily visible when simple estimation/calculation techniques are used. All articles in this special issue address how to minimise or adjust for coverage error. For NSOs, this raises resource use as another related question. Is the coverage issue the biggest one when considering AD? Should one accept no less than close to 100 percent coverage before even considering AD, or can one settle for less and combine AD with sampling techniques and modelling? Although essentially an estimation problem, censuses seem to have a 100 percent coverage target. While this is hard to achieve, it seems reasonable for legislative reasons and because of the census's importance for other social surveys. The trend seems to be that a combination of a traditional area-based frame field collection and AD sources is the choice for achieving this target. The AD source can compensate for undercoverage in field collection if the same people who are hard to reach are present in AD (e.g., through welfare-seeking systems). Alternatively, a field collection can be used in areas where it is believed the AD is poor.

In the business sphere there is a trade-off and often a good reason not to aim for 100 percent unit coverage. This is certainly the case in developing countries, but also applies elsewhere; it would be costly to keep the coverage of small home-based 'household' businesses up to date. To compensate for the undercoverage, other methods involving modelling and household surveys are needed.

3.4. Administrative Data in Developing Countries

In developing countries, the AD systems' maintenance and the contents coordination of planned and made investments are big barriers to using AD for statistics. Coverage issues are a result of these problems, not just in the sources themselves, but also in area frame-based sample surveys and in census practices. The ties between undertakings in population, agricultural and economic censuses and national AD systems are often weak or just occasional. The optimistic view of this is that developing countries may take advantage of 'leapfrogging' and develop AD system structures that facilitate standardisation and multiple uses (including statistics) from the start. Another good point is that a lot can be improved with relatively small means. To give two examples (among many): the first would be to add and enable geocodes in AD, censuses, and surveys. Adding coordinates to units (such as villages) in official databases would greatly improve the quality and simplify the updating of sampling frames and linking possibilities. Lack of harmonisation adds unnecessary burden in studies that combine several sources (e.g., Haslett et al. 2013). Simple actions like creating a standard geocode option for linking would free up analytic resources tied up in data cleaning. The second example would be to establish a statistical business register. By separating the concept of a statistical business register from that of an administrative business register, one can apply methods that achieve better alignment with the needs of national accounts and economic statistics (Wallgren and Wallgren 2007). The actions needed are country specific, but there are good and generic principles to follow. The African Development Bank's report

(ADB 2014) provides relatively exhaustive guidelines for a statistical business register. The guidelines are applicable outside Africa.

Although this special issue does not explicitly refer to developing country problems, the articles are still relevant, as coverage error is the of most concern statistical problem with AD. Some of the articles might be too advanced, but local NSO experts together with external consultants can benefit. In particular, the articles by Blackwell et al. and Burger et al. are good examples for countries such as those in Communidad Andina in South America. These countries have a number of AD sources for land, people, and business already in place and they are working on structures to use them for statistics.

4. A Final Note

Finally, I would like to congratulate the authors, guest editors, and the editors of JOS. Although there is plenty of literature about statistics and AD, a lot of it lacks the rigour that follows from a journal review process. A themed issue on administrative data is timely. With census transformation projects as a major driver, and as the area progresses further with theory meeting practice and vice versa, the future is likely to see a higher proportion of articles about AD methodology. It is an elusive thought (sometimes nursed at NSOs) that statistics based on AD is less complex. Because of society's growing appetite for data, methodologists are looking more closely at previously overlooked areas, and as I stress again the need to integrate data, many questions still need to be answered.

I predict a big increase of papers about AD and statistics, especially studies on the magnitude of error. Are the coverage errors large compared with other error types? What is the difference between having one controlled survey and having multiple data sources, and how is it addressed in terms of total survey design? Which configuration of data sources is the best? Are we sometimes making things worse when data sources are combined?

In particular, I believe the measurement properties of AD compared with those of survey data will be scrutinised. Several studies may conclude that direct collection and sample surveys are needed to adjust and/or guarantee statistical quality. Hence, even if there is a change in paradigm with the death of the sample survey as the *first* option to acquire data, this does not mean that a sample survey sometimes may not be the best option. Regardless of source configuration between AD and self-collected data, the choice depends on trade-offs between error types and cost. From a methodology perspective, it makes sense to further bridge the gap between the survey sampling tradition and the use of AD methods. In doing this, I do not believe that it helps to claim that the methodology applied to AD is a completely distinct and new area. Instead, a holistic view of the survey process is better, identifying where the methodological focus should lie, and when old methods are applicable or new ones need to be developed. NSOs that align their work with methods and their production environment with survey designs backing integrated data are better insured for the future.

5. References

African Development Bank. 2014. *Guidelines for Building Statistical Business Registers in Africa – Laying the Foundations for Harmonization of Economic Statistics Program*. Tunisia: Statistical capacity building division, Statistics Department

African Development Bank. Available at: http://www.afdb.org/en/knowledge/publications/guidelines-for-building-statistical-business-registers-in-africa/ (accessed 14 July 2014).

Andersson, C., A. Holmberg, I. Jansson, K. Lindgren, and P. Werner. 2013. "Methodological Experiences from a Register-Based Census." 2013 Joint Statistical Meetings, Montreal, 3–8 August 2013. http://www.amstat.org/sections/SRMS/Proceedings/y2013/Files/309549_82867.pdf (accessed July, 2015).

Griffin, R. 2014. "Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020." *Journal of Official Statistics* 30: 177–189. Doi: http://dx.doi.org/10.2478/jos-2014-0012.

Haslett, S., G. Jones, and A. Sefton. 2013. *Small-Area Estimation of Poverty and Malnutrition in Cambodia*. National Institute of Statistics, Ministry of Planning, Royal Government of Cambodia and the United Nations World Food Program. Available at: http://www.wfp.org/content/cambodia-small-area-estimation-poverty-and-malnutrition-april-2013 (accessed 14 July 2014).

Heath, T. and J. Goodwin. 2011. "Linking Geographical Data for Government and Consumer Applications." In *Linking Government Data*, edited by D. Wood, 73–92. New York: Springer.

Jäder, A. and A. Holmberg. 2005. *The Growth Rate Method for Production of Rapid Estimates of the Swedish Foreign Trade. Background facts on Economic Statistics* 2005:9. Statistics Sweden. Available at: http://www.scb.se/statistik/_publikationer/OV9999_2005A01_BR_X100ST0509.pdf (accessed 28 July 2014).

Mayer-Schönberger, V. and K. Cukier. 2013. *Big Data – A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Nordbotten, S. 1966. "A Statistical File System." *Statistisk Tidskrift* 4: 99–103. Available at: http://www.nordbotten.com/frame.html (accessed 14 July 2014).

Office of National Statistics (ONS). 2013. *Beyond 2011: Producing Population Estimates Using Administrative Data*. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html (accessed 14 July 2014).

Sadinle, M. and S.E. Fienberg. 2013. "Generalized Fellegi–Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." *Journal of the American Statistical Association* 108: 385–397. Doi: http://dx.doi.org/10.1080/01621459.2012.757231.

Scheuren, F. 1999. "Administrative Records and Census Taking." *Survey Methodology* 25: 151–160.

Schulte-Nordholt, E., M. Hartgers, and R. Gircour. 2004. *The Dutch Virtual Census of 2001. Analysis and methodology*. Voorburg/Heerlen. Available at: http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/publicaties/archief/2005/2001-b57-e-pub.htm (accessed 14 July 2014).

Skinner, C., J. Hollis, and M. Murphy. 2013. Beyond 2011*: Independent Review of Methodology*. Office of National Statistics. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011–independent-review-of-methodolgy/index.html (accessed 21 July 2014).

United Nations Economic Commission for Europe (UNECE). 2007. *Register-Based Statistics in the Nordic Countries. Review of Best Practices With Focus on Population and Social Statistics*. United Nations Publication. Available at: http://www.unece.org/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed 14 July 2014).

United Nations Economic Commission for Europe (UNECE). 2011. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. United Nations Publication. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed 14 July 2014).

Zaletel, M. and I. Krizman. 2008. "The Hidden Side of a Successful Story – Implication of Wide Use of Administrative Data Sources at National Statistical Institutes." Paper presented at the IAOS conference, Shanghai 14–16 October 2008.

Zhang, L.C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Wallgren, A. and B. Wallgren. 2007. *Register-Based Statistics Administrative Data for Statistical Purposes*. Chichester: John Wiley & Sons.

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 527–535, http://dx.doi.org/10.1515/JOS-2015-0032

Discussion

Stephen E. Fienberg[1]

1. Introduction

In countries around the world, support for traditional population censuses and large-scale sample surveys has eroded. The issues raised most often are cost, public distrust (privacy), and timeliness of data releases, and these are weighed against accuracy and alternative sources of population information (see e.g., Fienberg and Prewitt 2010). Official statisticians continue to explore alternatives and supplements to regular censal canvasses of the population and large-scale surveys for various purposes. Although various forms of population registers have replaced traditional head counts in a number of countries, new realities have led statisticians to question their accuracy and completeness. In some instances, political concerns present a problematic overlay to all of these developments.

So what has changed? First response rates in large-scale probability surveys have steadily declined, especially as individuals have become more mobile and in some senses less accessible to traditional methodologies, such as door-to-door person interviewing, telephone interviewing, and mail surveys. Second, family structures have also changed in ways that make households less apt as a unit of measurement. Third, migration, especially migration involving illegal immigrants, now poses a vexing problem in countries that have long professed not to be afflicted with such complexities, especially in the European Union. Finally, politicians have become increasingly suspicious of the costs of official statistics methodologies, believing that equivalent information could have been obtained from the internet at a fraction of the cost. Understanding the provenance and quality of the data has been the hallmark of official statistics as a field, but explaining the importance of this in the political climate of budget cuts and austerity is often difficult.

What has been the response to this situation by the official statistics community? Three mechanisms stand out:

(1) The use of administrative records, not solely but in combination with one another and more traditional census methods.
(2) The use of post-enumeration surveys to replace costly nonresponse follow up to such mechanisms involving telephone and personal interviewing,
(3) The use of online census forms.

[1] Maurice Falk University Professor of Statistics and Social Science at the Department of Statistics, the Heinz College, and the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A. Email: fienberg@stat.cmu.edu

Acknowledgment: The preparation of this discussion was partially supported by the U.S. National Science Foundation and the U.S. Census Bureau through NSF grant SES1130706 to the Department of Statistics, Carnegie Mellon University.

There are many hurdles, both methodological and practical, that need to be overcome before these approaches can yield the high-quality data that many government statistical agencies are recognized as producing. A major methodological issue is how to combine the potentially noisy information from multiple sources. In the United States, as we approach our next decennial census in 2020 the focus is on (1) and (3) with little focus on combining multiple sources for actual population estimation (see Eddy and Fienberg 2014), largely because of past political battles over post-enumeration surveys and sampling; for example, see Anderson and Fienberg (2001), Anderson et al. (2000), and Brown et al. (1999). In other countries the mix is different. Canada, for example, has pioneered the use of online forms, achieving a 54% online response, although set in a traditional mail-out mail-back context with mailed postcards replacing the initial mailing of census forms. But Canada, too, is exploring the use of administrative records going forward, see the overview in Hamel and Béland (2013). And in Israel, two separate surveys are used to capture the effects of both over- and underenumeration; see Nirel and Glickman (2009) for a detailed description.

2. The Articles in this Issue

The present issue of the *Journal of Official Statistics* includes contributions from statisticians in multiple countries, describing their current and recent efforts to reshape their census and population-reporting methodologies as well as more technical methodological contributions. Below I comment on some of the features of their efforts and both the novelties and the commonalities.

The primary methodological features that tie many of these articles together are (a) record-linkage methodologies, (b) the use of capture-recapture (dual systems) models and their multidimensional analogues when three or more record systems are involved, and (c) other related log-linear model methods. What I found especially gratifying in reviewing these contributions was the extent to which virtually every article linked in some way to my past and current research interests and activities on these topics. I mention a number of these links to my work as well as that of others in what follows.

2.1. Gerritse, van der Heijden, and Bakker

These authors explore the departure from independence in the dual-systems approach to two registers by modeling different values of the dependence. They do this by looking at a range of "offset" values and by considering the role of covariates, albeit with missing values. They use data from Dutch registers and introduce a stratification by country of nationality that allows for some common modeling of dependencies. They briefly discuss extensions to multiple registers based on standard multiple-recapture approaches, first laid out in Fienberg (1972) and Bishop et al. (1975), and something they refer to as the multiplier method, which presumes independence of sources as well as the availability of some joint information between them. The latter does not generalize in any useful way for most census problems I know. An alternative approach to this work worthy of note is that presented by Kurtz (2014) who uses a localized approach to adjust for interactions with covariates.

2.2. Zhang

The focus in this article is on adjusting census results for both underenumeration (omissions) using coverage survey data and overenumeration (erroneous enumerations). The U.S. Census Bureau approach is to do this by directly adjusting for overenumeration for a sample of census blocks prior to the use of dual-systems methods for omissions; for example, see Hogan (1993). There are two strong assumptions involved: (1) that all erroneous enumerations can be identified in the census data for the sample blocks, and (2) that there are no erroneous enumerations in the coverage survey for those blocks. The Israeli Central Bureau of Statistics approach uses two separate surveys, one for omissions and another to remove erroneous enumerations; for example, see Nirel and Glickman (2009). Through a targeted class of models and assumptions regarding the two surveys they estimate the population totals, adjusting for both kinds of errors. Zhang adopts a somewhat different strategy involving two coverage surveys, or a coverage survey and an administrative list, and he directly models the resulting three-dimensional cross classification using log-linear models, and a variant on them also involving marginal log-linear models. By assuming that the second coverage survey only has underenumerations (and no matching error), he identifies a model that allows for estimation of the relevant population totals. He approaches estimation through a form of moment estimation. I would have preferred a maximum-likelihood approach as that would have directly allowed for consistency and asymptotic variances.

The earlier literature on multiple-system estimation (with three or more lists) includes many ways to estimate population totals using log-linear and related models, which allow for both dependence among lists and individual-level heterogeneity, but they all make the unrealistic assumption of no overenumeration (or at least that overenumeration can be corrected without error). See Darroch et al. (1993), Fienberg et al. (1999), Manrique-Vallier and Fienberg (2008), and the International Working Group for Disease Monitoring and Forecasting (1995a, 1995b) for some examples. Zhang's approach, moved to a setting with at least four lists, might allow for a weakening of this assumption.

2.3. Di Consiglio and Tuoto

There are two possible ways of dealing with the uncertainty associated with record linkage for various lists, especially in a censal context. We can actually examine the error process in linkage and propagate it through into the population estimation process (see below for more details) or we can develop simplified models involving homogeneous probabilities of true and false links and then either estimate those probabilities using auxiliary data or explore the sensitivity of the assumption of no linkage error by considering a range of possible values for the parameters. Ding and Fienberg (1994, 1996) took the latter approach. In this article, Consiglio and Tuoto take up the Ding and Fienberg approach for two lists (1994) and extend it to allow for a slightly more complex possibility of true and false probabilities of links. Adapting their approach to multiple lists and exploring how well the approach captures the effect of heterogeneous errors of linkage associated with basic record linkage approaches are natural next steps. For related literature borrowing from methods used in the context of tag loss in animal population studies, see Seber et al. (2000) for the two-list case and Lee et al. (2001) for an extension to multiple lists.

2.4. Yildiz and Smith

These authors address a related but different problem involving the combination of administrative records data than those looking to create and analyze individual-level data, which are ultimately aggregated into the form of a contingency table. Rather, they consider adjusting one source of aggregate administrative data in the form of a cross classification to conform with "more accurate" values from an auxiliary dataset. The methodology is once again based on log-linear models, where the use of "offsets" allows the preservation of interaction structures from the original aggregate administrative data. They implement the approach using maximum-likelihood estimation based on iterative proportional fitting as described in Bishop et al. (1975).

What if there are multiple updated data sources for overlapping variables? This turns out to be a far more complex problem, and requiring consistency of overlapping margins for the updated sources is not sufficient to ensure the existence of a contingency table with the original interaction structure and adjusted margins to accord with the updated sources. For a related discussion of nonexistence in the context of privacy protection, see, for example Yang et al. (2012).

2.5. Chipperfield and Chambers

As I and others have noted (e.g., see Bilenko et al. 2003), most methods for record linkage, and especially those used in the context of official statistics, are based on variants of the now classic methodology of Fellegi and Sunter (1969). The book by Herzog et al. (2007) provides a useful update including suggested links to multiple-recapture methodology. It also describes the use of the FS methodology for duplicate detection as well as record linkage, and in practice these are done as separate activities rather than jointly.

This article is set in this FS record-linkage tradition and focuses on the use of bootstrap methodology to propagate the uncertainty from the matching process of two files into the subsequent statistical analyses. It extends the authors' earlier work in which the results of the record linkage are used for regression modeling to the setting of linked categorical data, with a special focus on analyses based on log-linear or logistic models. They assume no duplicates within the two files to be used for linkage, and also that fields within records used for linkages either match or not. The latter is an impractical restriction for files involving alphanumerics (such as names and places) or even dates, and this can clearly be relaxed. They begin with the assumption that the files to be linked are of equal size and assume 1-1 matching, a case known to produce high probabilities of linkage, and then proceed to the more practical situation where some files will not match and the files will be of differing lengths. They demonstrate the methodology in a simulation and then in the context of the 2011 census of Australian Population and Housing, comparing it with two alternative procedures with interesting summary results. When it comes to the actual implementation of record-linkage methods, the devil is in the detail, and we clearly need more details to judge the utility of the methods presented here.

There is a hint buried in this article regarding the potential representation of linkages in terms of network structures. This is the approach adopted by Sadinle (2014) in the context of a duplicate detection task, and extended in Sadinle (2015) to joint duplicate detection and record linkage involving the possibility of multiple files. See also the related approach in Steorts et al. (2014a, 2014b). Getting these methods to scale so that they work in a censal context remains a challenge.

2.6. Blackwell, Charlesworth, and Rogers

These authors describe in considerable detail the way in which the U.K. Office of National Statistics used administrative records and record linkage in a sample within 58 carefully selected local areas corresponding to formal administrative authorities. They describe aspects of both over- and undercoverage. There is a serious technical issue here for anyone familiar with the actual uses of Fellegi-Sunter methodology at the scale of census files, even for restricted geographical areas. As the size of lists to be matched grows, the number of possible matches increases as the product of their sizes. The only way around this is through blocking, but this rests on the assumption that the blocking variable is error free, something that is rarely the case. It will not come as a surprise that the effort was complex. At any rate, the authors do not provide any specific details on the implementation. Given my understanding of the approach used in the U.S. for coverage evaluation, the actual implementation rarely resembles what one finds in research papers and the published literature. Even without the details, their presentation here should serve as a sobering reminder that we require much more than elegant theory and methods to produce official statistics of value.

2.7. Bryant and Graham

These authors have as their goal replacing the traditional census in New Zealand with an administrative census, perhaps supplemented by a coverage survey, to (i) generate population estimates, and (ii) assemble individual-level socioeconomic data for different data sources. Their key methodological focus is on the use of Bayesian hierarchical models, something I approve of heartily (cf. Fienberg 2011 and Fienberg et al. 1999). The unique feature of the present article, in the context of this special issue on coverage measurement, is its reliance on a demographic accounting framework. By working with multiple data sources and using the hierarchical structure to "borrow strength" across sources via the hierarchical structure, they explain how to properly account for random variation in the demographic series and in the measurement of these series and how to propagate this uncertainty into final population estimates. Accuracy of immigration and emigration numbers is crucial. However, what remains unclear from their article is how they will ultimately combine multiple sources and a coverage survey to provide individual-level data. Nor do they discuss how they propose to account for differences in the timing of data sources and target populations.

2.8. Burger, van Delden, and Scholtus

These authors present an approach to estimate the bias and variance resulting from classification errors in a business register. Their approach involves a homogeneous misclassification model, based on a two-parameter transition matrix, and bootstrap resampling methodology. They then explore the sensitivity of bias and variance specification of the transition matrix via a simulation study. Their study focuses on a single data source, and what would be of special interest in the context of the problems explored in the other articles in this issue is how the kinds of misclassification errors would affect record linkage of multiple sources and statistical calculations on the resulting merged data.

3. Conclusion

Some areas that still require research to bring the theory and methods described above into actual practice in official statistics settings are as follows:

- *Record linkage methods from three or more files*. As I note above, virtually all of the record-linkage methodologies in use today deal only with pairs of files, for example, from a census and a coverage survey. But as we move to the use of administrative records, we need to think in terms of multiple files, and it does not suffice to match each separately into some common file. Our census-related project at Carnegie Mellon has produced a number of efforts in this direction, see Sadinle (2014, 2015), Sadinle et al. (2011), Sadinle and Fienberg (2013), Steorts et al. (2014a, 2014b) and Ventura et al. (2014). See also Fienberg and Manrique (2009). However, making these methods suitable for "industrial strength" applications such as those involved in census production systems will take considerable time and effort.
- *Combining duplicate detection and record linkage*. Many agencies and researchers treat duplicate detection within files and record linkage between files as separate activities. They are not and the very fact that the Fellegi-Sunter is used for both should get us thinking about ways to approach the tasks in a unified way. Sadinle (2014, 2015) and Steorts et al. (2014a, 2014b) adopt such unifying approaches, and the Bayesian framework that they use is especially suitable for this. The type of combining raises even greater issues of scalability than methods for record linkage alone.
- *Propagating duplicate detection and record linkage error into subsequent calculations.* Researchers have recognized this problem for decades, but it has received renewed attention over the past decade. The article by Chipperfield and Chambers described above and the related ones by these authors take a frequentist approach and develop bootstrap procedures to accompany them. Sadinle (2014, 2015) and Steorts et al. (2014a, 2014b), following earlier work of Belin and Rubin (1995), Larsen and Rubin (2001), and Tancredi and Liseo (2011), focus on Bayesian approaches. Implementing any of these methods at scale is complex.
- *Measuring both EEs and omissions.* Much of the discussion surrounding the use of coverage surveys and capture-recapture methods for population estimation focuses on the problem of omissions and census undercount. But erroneous enumerations often play as big a role, and will take on increasing importance as administrative record systems often contain many records that are out of scope. The approach of Nirel and Glickman (2009) suggests one way of proceeding, but with attendant assumptions, and that of Zhang suggests that simplifying assumptions and models will be useful in other contexts. Much more work needs to be done here.
- *Putting it all together with statistical models and methods.* The propagation of error issue discussed above is but one of the multiple aspects of a unified methodology for census taking. We have no such overarching framework and methodology today.

- *Moving from new methods to statistical practice.* For the new approaches to census taking to work in practice, agencies will need to be able to address essentially all of the problems listed in the bullets above, and combine them with effective use of online forms and real-time editing. In the U.S, Canada, and the U.K., as well as in other countries, censuses have combined the study of housing, households, and individuals. Most of my discussion has focused on the latter, but placing individuals in households also requires special attention, as does the linkage of households across multiple files.

While most of the articles in this special issue of the *Journal of Official Statistics* focus primarily on coverage issues in censuses and administrative record systems, many of the concerns and methodologies are relevant to the future of large-scale sample surveys as well. The changes demanded by government officials and the public are great and likely to continue increasing. The good news in the work reported on here and in other recent methodological developments is that we may be entering a new era for collaboration between statistical offices and academic statisticians.

4. References

Anderson, M.J., B.O. Daponte, S.E. Fienberg, J.B. Kadane, B.D. Spencer, and D. Steffey. 2000. “Sample-Based Adjustment of the 2000 Census – A Balanced Perspective.” *Jurimetrics* 40: 341–356.

Anderson, M.J. and S.E. Fienberg. 2001. *Who Counts? The Politics of Census-Taking in Contemporary America*. Rev. ed. New York: Russell Sage Foundation.

Belin, T.R. and D.B. Rubin. 1995. “A Method for Calibrating False-Match Rates in Record Linkage.” *Journal of the American Statistical Association* 90: 694–707. Doi: http://dx.doi.org/10.1080/01621459.1995.10476563.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, and S.E. Fienberg. 2003. “Adaptive Name Matching in Information Integration.” *IEEE Intelligent Systems* 5: 16–23. Doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2003.1234765.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. Reprint, New York: Springer, 2007.

Brown, L.D., M.L. Eaton, D.A. Freedman, S.P. Klein, R.A. Olshen, K.W. Wachter, M.T. Wells, and D. Ylvisaker. 1999. “Statistical Controversies in Census 2000.” *Jurimetrics* 39: 347–376.

Darroch, J., S.E. Fienberg, G. Glonek, and B. Junker. 1993. “A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability.” *Journal of the American Statistical Association* 88: 1137–1148. Doi: http://dx.doi.org/10.1080/01621459.1993.10476387.

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

Ding, Y. and S.E. Fienberg. 1996. “Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors.” *Survey Methodology* 22: 55–64.

Eddy, W.F. and S.E. Fienberg. 2014. *Envisioning the 2030 U.S. Census*. Morris Hansen Lecture, Washington Statistical Society, in preparation.

Fellegi, I. and A. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64: 1183–2010. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 409–439. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Fienberg, S.E. 2011. "Bayesian Models and Methods in Public Policy and Government Settings." *Statistical Science* 26: 212–226. Doi: http://dx.doi.org/10.1214/10-STS331.

Fienberg, S.E., M. Johnson, and B. Junker. 1999. "Classical Multilevel and Bayesian approaches to Population Size estimation Using Multiple Lists." *Journal of the Royal Statistical Society. Series A* 162: 383–406. Doi: http://dx.doi.org/10.1111/1467-985X.00143.

Fienberg, S.E. and D. Manrique-Vallier. 2009. "Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation." *Advances in Statistical Analysis* 93: 49–60.

Fienberg, S.E. and K. Prewitt. 2010. "Save Your Census." *Nature* 466: 1043, August 26, 2010. Doi: http://dx.doi.org/10.1038/4661043a.

Hamel, M. and Y. Béland. 2013. "Future Developments on the Canadian Census of Population." In Proceedings of the ISI World Statistical Congress, August 25-30, 2013, Hong Kong. Available at: http://www.statistics.gov.hk/wsc/IPS094-P3-S.pdf (accessed August 6, 2015).

Herzog, T., F. Scheuren, and W. Winkler. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer-Verlag.

Hogan, H. 1993. "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060. Doi: http://dx.doi.org/10.1080/01621459.1993.10476374.

International Working Group for Disease Monitoring and Forecasting. 1995a. "Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058.

International Working Group for Disease Monitoring and Forecasting. 1995b. "Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Human Diseases." *American Journal of Epidemiology* 142: 1059–1068.

Kurtz, Z. 2014. "Smooth Post-Stratification for Multiple Capture-Recapture." Unpublished manuscript. Available at: http://arxiv.org/abs/1302.0890. (accessed August 6, 2015)

Larsen, M.D. and D.B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41. Doi: http://dx.doi.org/10.1198/016214501750332956.

Lee, A.J., G.A.F. Seber, J.K. Holden, and J.T. Huakau. 2001. "Capture-Recapture, Epidemiology, and List Mismatches: Several Lists." *Biometrics* 57: 707–713. Doi: http://dx.doi.org/10.1111/j.0006-341X.2001.00707.x.

Manrique-Vallier, D. and S.E. Fienberg. 2008. "Population Size Estimation Using Individual Level Mixture Models." *Biometrical Journal* 50: 1051–1063. Doi: http://dx.doi.org/10.1002/bimj.200810448.

Nirel, R. and H. Glickman. 2009. “Sample Surveys and Censuses.” In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, 539–565.

Sadinle, M. 2014. “Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach.” *Annals of Applied Statistics* 8: 2404–2434. Doi: http://dx.doi.org/10.1214/14-AOAS779.

Sadinle, M. 2015. “A Bayesian Partitioning Approach to Duplicate Detection and Record Linkage.” Unpublished Ph.D. Dissertation. Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.

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. Doi: http://dx.doi.org/10.1080/01621459.2012.757231.

Sadinle, M., R. Hall, and S.E. Fienberg. 2011. “Approaches to Multiple Record Linkage.” In Proceedings of the ISI World Statistical Congress, 21-26 August 2011, Dublin, 1064–1071.

Seber, G.A.F., J.T. Huakau, and D. Simmons. 2000. “Capture-Recapture, Epidemiology and List Mismatches: Two Lists.” *Biometrics* 56: 1227–1232.

Steorts, R., R. Hall, and S.E. Fienberg. 2014a. “SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication.” *Journal of Machine Learning Research* 33: 922–930. Available at: http://jmlr.csail.mit.edu/proceedings/papers/v33/steorts14.pdf (accessed August 6, 2015).

Steorts, R., R. Hall, and S.E. Fienberg. 2014b. “A Bayesian Approach to Graphical Record Linkage and De-duplication.” Under Submission. Available at: http://arXiv:1312.4645

Tancredi, A. and B. Liseo. 2011. “A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems.” *Annals of Applied Statistics* 5: 1553–1585. Doi: http://dx.doi.org/10.1214/10-AOAS447.

Ventura, S. and R. Nugent. 2014. “Hierarchical Clustering with Distributions of Distances for Large-Scale Record Linkage.” In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, pp. 283–298. Berlin: Springer Link. Lecture Notes in Computer Science, Vol. 8744. Doi: http://dx.doi.org/10.1007/978-3-319-11257-222.

Yang, X., S.E. Fienberg, and A. Rinaldo. 2012. “Differential Privacy for Protecting Multidimensional Contingency Table Data: Extensions and Applications.” *Journal of Privacy and Confidentiality* 4. Available at: http://repository.cmu.edu/jpc/vol4/iss1/5 (accessed August 6, 2015).