ÖAW AUSTRIAN ACADEMY OF SCIENCES

VIENNA INSTITUTE OF DEMOGRAPHY

WORKING PAPERS

15/2017

ESTIMATING POPULATION COUNTS WITH CAPTURE-RECAPTURE MODELS IN THE CONTEXT OF ERRONEOUS RECORDS IN LINKED ADMINISTRATIVE DATA

DILEK YILDIZ, PETER G.M. VAN DER HEIJDEN AND PETER W.F. SMITH

Vienna Institute of Demography
Austrian Academy of Sciences
Welthandelsplatz 2, Level 2 | 1020 Wien, Österreich
vid@oeaw.ac.at | www.oeaw.ac.at/vid

Vienna Institute of Demography

Abstract

In the absence of a traditional census and a comprehensive population register (as it is the case in the UK), administrative data sources, i.e. health, school or tax records, can offer an alternative to estimate the size of residence population. However, such data sources are designed to capture information only from specific populations which imposes a challenge to the estimation. A suitable method to overcome the challenge is to link administrative data sources that collect information from different but overlapping populations and use capture-recapture models to estimate the population counts. There are various assumptions required to obtain unbiased estimates by using capture-recapture models. In practice, especially the assumptions on 'homogeneous inclusion probabilities' and 'no over coverage' are often not met. This paper proposes a two-step procedure for estimating population counts with capture-recapture models that account for heterogeneity of inclusion probabilities and the over coverage in the data sources. We apply our methodology to the linked Patient Register and Customer Information System dataset which violates both of the aforementioned assumptions. The Patient Register includes people who are registered with a National Health Service General Practitioner doctor. In 2011, the Patient Register overestimated the size of England and Wales population by 4.3% (over coverage) and its sex ratio was different than the 2011 Census estimates (heterogeneous of inclusion probabilities). The Customer Information System dataset provides information on all individuals who have ever had a national insurance number and children whose parents have made a child benefit claim relating to them. In 2011, it over estimated the size of England and Wales population by 9.5% and its age and sex structure was different than the 2011 Census estimates. Applying our approach, we estimate population counts of the South East region of England by age, sex and local authority, and compare them with census estimates using percentage difference maps.

Keywords

Administrative data, capture-recapture models, combining data, log-linear model with offset, population estimates, dual-system estimation.

Authors

Dilek Yildiz, (corresponding author), Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Vienna Institute of Demography/Austrian Academy of Sciences, International Institute for Applied Systems Analysis, Austria.
Email: dilek.yildiz@oeaw.ac.at

Peter G.M. van der Heijden, Utrecht University, Social Sciences, The Netherlands and University of Southampton, Southampton Statistical Sciences Research Institute, United Kingdom.
Email: P.G.M.vanderHeijden@uu.nl

Peter W.F. Smith, University of Southampton, Administrative Data Research Centre for England, United Kingdom.
Email: P.W.Smith@soton.ac.uk

Acknowledgements

This research is based on the PhD thesis of Dilek Yildiz which was funded jointly by Economic and Social Research Council and Office for National Statistics.

Estimating Population Counts with Capture-Recapture Models in the Context of Erroneous Records in Linked Administrative Data

Dilek Yildiz, Peter G.M. van der Heijden, Peter W.F. Smith

1 Introduction

In the absence of a traditional census, an alternative method to estimate population counts and other census variables is to use individual linkage of existing administrative sources/registers. Nordic countries have been estimating their population counts by register-based censuses since the 1980s (Statistics Finland, 2004). They have specific population registers that aim to capture the whole population. When it is not possible to obtain population estimates from one such register directly, it is possible to use data from two or more registers. This requires linking registers which are not necessarily designed to capture the usual resident population[1] as aimed by traditional censuses. In this case, it is possible that some of the usual residents are not included in either of the linked administrative data sources, and will be missing in the final population estimate, leading to underestimation of the population size. Moreover, it is also possible that some of the people registered in the administrative data sources may not be eligible to be included in the usual resident population, and therefore may lead to overestimation of the population size.

It is well documented that underestimation of population sizes in linked data sources can be dealt with using the capture-recapture approach also known as dual-system and dual-record system, or multiple-recapture, multiple-system, and multiple-records system (IWGDMF, 1995; Bishop *et al.*, 1975). It is used both in countries which are conducting traditional censuses to estimate under coverage after post enumeration surveys and in countries which estimate their population by register-based censuses. In the latter case, a second register is used as the recapture sample instead of a post enumeration survey.

There are five assumptions required to obtain unbiased estimates by using capture-recapture models (IWGDMF, 1995; Bishop *et al.*, 1975) as follows:

- Probability of being captured in one source is independent of probability of being captured in the second source

[1]For the census purposes, "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).

- The probability of being captured or recaptured is homogeneous across individuals
- The population is closed
- The individual linkage between the sources are perfect, i.e. no linkage error
- There is no-over coverage in the source

When these assumptions are violated, the estimates will be biased. Especially when the capture-recapture approach is used to estimate population counts by using administrative registers, it is common that at least one of these assumptions will not hold. Hence, a number of studies have been investigating how to avoid or reduce the bias in the estimates when at least one of the assumptions is violated (Cormack, 1989; Darroch *et al.*, 1993; Brown *et al.*, 2011; Van der Heijden *et al.*, 2012; Gerritse *et al.*, 2015; Zhang, 2015). Likewise, this paper deals with the violation of the "no over coverage" assumption.

This paper contributes to the literature on estimating population counts using registers which are not designed to record the usual resident population. For this purpose traditional log-linear capture-recapture models are extended to cope with both under- and over-coverage. To illustrate the methodology, we use the Patient Register and the Customer Information System (CIS) to estimate the usual resident population of the South East region of England. The Patient Register and the CIS are two comprehensive administrative data sources in England and Wales which collect information from different, but overlapping population groups. As discussed in ONS (2012, 2013a) they are subject to both underestimation and overestimation. The Patient Register includes people who are registered with a National Health Service General Practitioner doctor. In 2011, the Patient Register overestimated the England and Wales population by 4.3% (over coverage) and its sex ratio was different than the 2011 Census estimates (heterogeneous of inclusion probabilities). The Customer Information System dataset provides information on all individuals who have ever had a national insurance number and children whose parents have made a child benefit claim relating to them. In 2011, it over estimated the general population by 9.5% and its age and sex structure was different than the 2011 Census estimates. Hence, the aim of this paper is to develop a methodology to estimate the number of people in a particular age group, sex and local authority by using information from the individually linked Patient Register and CIS.

The capture-recapture approach is used to estimate the number of people who are usual residents, but who are neither recorded in the Patient Register nor in the CIS. However, since both the Patient Register and the CIS exceed the census estimates (ONS, 2012, 2013a), it is inevitable that the capture-recapture approach will also overestimate the population total by adding an estimated count for people who are not registered in both datasets. Even though the capture recapture approach increases the overestimation in the population size, the pre-

dicted values from the capture-recapture models present the age group, sex and local authority association structure in the population more accurately than the direct estimates from these sources since they involve people who are not registered with either of the administrative sources. The association structures improve even more when they are estimated by the extended capture-recapture models which take the heterogeneity of inclusion probabilities over levels of age, sex and local authority into account. To adjust the overestimation, we use the log-linear models with offsets to combine the linked Patient Register and CIS information with auxiliary marginal information. Yildiz & Smith (2015) combined auxiliary marginal information with population counts from the Patient Register. In the second step of the estimation, this paper follows their work, and uses the marginal age group, sex and age group-sex tables from the 2011 Census estimates as the auxiliary source. However, in the future it is possible to take the marginal tables from another source, such as another administrative source, an annual survey, or a coverage survey. Moreover, in the first step of the estimation, we improve on Yildiz & Smith (2015) and combine auxiliary marginal information with population counts estimated by different capture-recapture models instead of using population counts from one administrative source.

The paper has been divided into five sections, including this section. Section 2 deals with the methodology used to estimate population counts. Section 3 presents the model specification and the results. Finally, the last section gives a brief summary and discussion of the findings.

2 Method

The linked dataset provides information about sizes of three population groups: those who are only registered with the Patient Register, those who are only registered with the CIS, and those who are registered with both the Patient Register and the CIS. Although, both of the linked registers collect information from the majority of the population, estimating the size of the usual resident population directly by using the linked dataset is problematic in two ways:

- People who are not registered with any of these registers, but are usual residents of South East region lead to underestimation of the true population size.
- People who are not usual residents of the South East region, but registered with at least one of the registers lead to overestimation of the true population size.

This paper deals with the underestimation and the overestimation in two steps: 1) log-linear capture-recapture models and 2) log-linear models with offsets. The estimates from the first

step models are used as offsets in the second step. Both of the steps include three options as shown in Table 1.

1) Log-linear capture-recapture models	2) Log-linear models with offsets (Offsets are estimates from the first step)
a) People registered in either sources	a) A Model
b) Classic capture-recapture model	b) A,S Model
c) Parsimonious capture-recapture models	c) AS Model

Table 1: Estimation steps

The first step, the capture-recapture approach, is used to estimate the number of people who are usual residents but are neither recorded in the Patient Register nor in the CIS. This approach has also proved to be useful to correct the heterogeneous inclusion probabilities across age groups, sex and local authorities. However, because both registers include people who are not defined as usual residents, hence overestimate the total population, the estimates from the first step, capture-recapture approach, also overestimate the population size. Therefore, in a second step, log-linear models with offsets are used to adjust the predicted values from the capture-recapture models by combining them with auxiliary information. Accordingly, this section continues with explaining how the capture-recapture approach and the log-linear models with offsets are used in the context of estimating population counts from two linked registers which are subject to both underestimation as well as overestimation. Similar log-linear models with offsets are discussed in detail in Yildiz & Smith (2015). Hence, this section focuses more on the capture-recapture models.

In the first step we consider three options to estimate the complete population size by using information from the linked Patient Register and CIS dataset which are listed as follows:

a The linked dataset is neither subject to under coverage or over coverage. The true usual resident population size, N, is equal to the population captured by either of these two registers, n.

b The registers do not capture the complete usual resident population but it is possible to estimate the under coverage by the classic capture-recapture approach.

c The registers do not capture the complete usual resident population or satisfy the assumptions for the classic capture-recapture approach. For this option we use parsimonious log-linear models that take the heterogeneity of inclusion probabilities into account.

In the second step, we use log-linear models with offsets to adjust for overestimation of the predicted values estimated by these options. Recall that the offsets are the estimated

population counts from the first step. The second step also consists of three options depending on the auxiliary marginal information used to adjust the over estimation.

a Only the auxiliary age group distribution is combined with different offsets (A model).

b Both the auxiliary age group and sex distributions are combined with different offsets (A,S model).

c The auxiliary age group-sex marginal information is combined with different offsets (AS model).

The capture-recapture approach is used to estimate the sizes of closed populations i.e. populations where the size is fixed, and not affected by birth, death or migration. The size of the population is estimated by treating one sample as captured and a second sample as recaptured data. Table 2 provides an overview of the information available from a linked dataset where $\hat{x}_{00}$ refers to the missing part of the population. Traditionally, the capture-recapture approach has been used to estimate the sizes of animal populations in ecological research. Its use in estimating the size of human populations is more recent and mostly focused on epidemiology (IWGDMF, 1995). It is also used to estimate (socially) hard to count population sizes of illegal immigrants (Van der Heijden *et al.*, 2012). Another use of the capture-recapture approach is to estimate the coverage of a traditional census by using data collected by a post-enumeration survey as the recaptured sample. This approach is used in the United Kingdom, the United States, Australia, New Zealand, Turkey, and Switzerland (Chen & Tang, 2011).

	Second register		
First register	Yes	No	Total
Yes	x_{11}	x_{10}	x_{1+}
No	x_{01}	$\hat{x}_{00}$	
Total	x_{+1}		

Table 2: A capture-recapture table

The classic capture-recapture approach assumes that the inclusion in the first register is statistically independent of inclusion in the second register, and the probability of being captured or recaptured is homogeneous across individuals for at least one of the registers (Zwane *et al.*, 2004). However, these assumptions do not always hold in human populations. For example, people registered with one of the registers may be more likely to register with the other (social visibility) or some people may be less likely to register with both administrative registers (social invisibility).

The probability of being captured by these registers may change according to covariates, i.e. age groups, sex, and local authorities. In this case it is more convenient to use less restrictive assumptions, and fit conditional independence models (Bishop *et al.*, 1975; Van der Heijden

et al., 2012). Thus we deal with the violation of the homogeneous inclusion probabilities assumption by including interaction terms to the log-linear capture-recapture models, and fitting conditional independence models to a higher-way table instead of to a two-way table. These models produce less biased estimates of the age group-sex-local authority association structure in the population than the direct estimates from either a single register or the linked dataset. Inclusion of covariates also reduces the bias that may have been caused by the violation of independence of inclusion probabilities assumption by reducing the heterogeneity in the dataset (Gerritse *et al.*, 2015).

For this paper, the individual datasets were not available. Hence, it is assumed that the Patient Register and the CIS are perfectly linked. It is worth noting that the violation of perfect linkage assumption could lead both to under coverage and over coverage (Bakker & Daas, 2012).

Consider two registers (here the Patient Register and the CIS) that are linked and the data are presented in a 2×2 table such as in Table 2, where x_{1+}, x_{+1}, and x_{11} denote the observed sizes of the population in the first register, in the second register and their overlap respectively. The observed population sizes are denoted by x_{ij}, the expected population sizes are denoted by m_{ij}, and the fitted population sizes are denoted by $\widehat{m}_{ij}$. Assume the size of missing population is m_{00}, the unknown size of the total population is N, and the size of the observed population covered by both registers is

$$n = x_{11} + x_{10} + x_{01}. \tag{1}$$

In case of two registers, Bishop *et al.* (1975) presents two approaches to estimate the size of the population which are not recorded in either of the registers: the basic approach and the incomplete table approach. The basic approach is based on the probabilities of individuals being in the first, in the second and in both registers. This approach results in the well-known estimator for the total number of individuals in the population equal to

$$\widehat{N} = \frac{x_{1+}x_{+1}}{x_{11}}. \tag{2}$$

The second approach yields the maximum likelihood estimate of the missing cell equal to

$$\widehat{m}_{00} = \frac{x_{10}x_{01}}{x_{11}}, \tag{3}$$

$$\widehat{N} = n + \widehat{m}_{00}. \tag{4}$$

It is also possible to calculate the same $\widehat{N}$ by using the log-linear independence model. Models including up to three parameters can be fitted to the table for two registers since it has three

cells with observed counts. Accordingly, the log-linear independence model for a two-way table can be written as

$$\log m_{ij} = \lambda + \lambda_i^I + \lambda_j^J \tag{5}$$

where the row variable (the first register/the Patient Register) is denoted by I and the column variable (the second register/the CIS) is denoted by J where $\lambda_0^I = \lambda_0^J = 0$, $m_{00} = \exp(\lambda)$, and $m_{ij} = E(x_{ij})$. Because the number of parameters in the independence model is equal to the number of cells with the observed counts, it is also the saturated model and fits perfectly in the sense that the observed counts equal the fitted counts, $\widehat{m}_{ij} = x_{ij}$ if $(i,j) \neq (0,0)$. The capture-recapture approaches explained above and the saturated log-linear model provide the same population estimate $\widehat{N}$.

In this paper we used a five-way $I \times J \times A \times S \times L$ table where I and J are binary variables regarding being recorded in the Patient Register and in the CIS. A, S and L denote the five-year age groups, two genders and local authorities in the South East region respectively. The observed and the estimated population counts in each cell are denoted by x_{ijasl} and $\widehat{m}_{ijasl}$. Each model is denoted by a list of the highest-order terms fitted to the data. For example, the mutual independence model is denoted by the symbol I,J,A,S,L and the saturated model is denoted by IASL,JASL. In normal circumstances the saturated log-linear model for a five-way table would be the IJASL model. However, for the linked data the saturated model is the IASL,JASL model since its number of parameters is equal to the number of cells (recall that x_{00} cell is missing in each 2×2 table).

Once the $\widehat{m}_{++asl}$ cells are estimated by capture-recapture models, they are combined with auxiliary information to reduce the bias and produce the final population estimates. The combination takes place by using log-linear models with offsets. This allows us to preserve the age group-sex-local authority association of the estimates from Step 1 options while updating margins according to the auxiliary information.

Log-linear models with offsets have been used to combine information from different data sources to estimate migration (Raymer & Rogers, 2007; Raymer *et al.*, 2007, 2009, 2011; Smith *et al.*, 2010) and population counts (Yildiz & Smith, 2015). It has been shown that combining information from the Patient Register and the census estimates by using log-linear models with offsets provided accurate estimates for population counts.

This paper extends the approach taken in Yildiz & Smith (2015) by using more accurate age group-sex-local authority association structures developed by the capture-recapture models as offsets instead of using the association structures from the Patient Register directly. Another difference is the use of more parsimonious log-linear models with offsets. By using parsimonious models both the possibility of overfitting and the variance decrease (Agresti, 2013). We

refer to Agresti (2013, p.128); Raymer & Rogers (2007); Raymer *et al.* (2009); Smith *et al.* (2010); Yildiz & Smith (2015) for detailed explanation of log-linear models with offsets, and present the application for this study in the next section.

3 Estimating the Size of the South East Population

This section is divided into two subsections according to the offsets used in the second step. In the first subsection we evaluate models in which offsets are either option a) the linked data or option b) the classic capture-recapture approach. The second subsection presents the models in which the offsets are equal to the estimates from option c) parsimonious capture-recapture models.

3.1 The Linked Data and the Classic Capture-Recapture Approach

Offsets used in log-linear models in this section are equal to a) the observed counts in the linked data, and b) the estimates from the classical capture-recapture model separately. Then the estimates obtained from the two options are combined with auxiliary data by log-linear models with offsets to correct the overestimation as a second step.

A basic idea to estimate population counts from the linked data source is to assume that everyone in the usual resident population is at least registered with the Patient Register and/or the CIS, and only the usual residents can be registered with any of these sources (option a). In other words, the total unknown population count N is equal to the observed population count n in the linked data. Denote the first register in Table 2 as the Patient Register and the second register as the CIS. Assume each cell in the table presents the observed population for cell ij for a particular age group a, sex s, and local authority l. Then, it is possible to compute the population of the South East region by adding people who are registered with at least one of these two administrative sources. Hence, the population estimated for age group a, sex s, and local authority l is equal to

$$\widehat{m}_{++asl} = x_{11asl} + x_{10asl} + x_{01asl}, \tag{6}$$

and for simplicity we can collapse over i and j and denote by $\widehat{m}_{asl}$.

The left panel of Figure 1 presents the census estimates and population counts in the linked data, sex ratios of the census estimates and population counts of the linked data, and the mean percentage differences between the census estimates and population counts of the linked data for males and females by age groups, for the South East region. As expected, the linked

data over enumerates the population for almost all of the age groups and its sex ratio exceeds the census estimates sex ratio for most of the age groups.

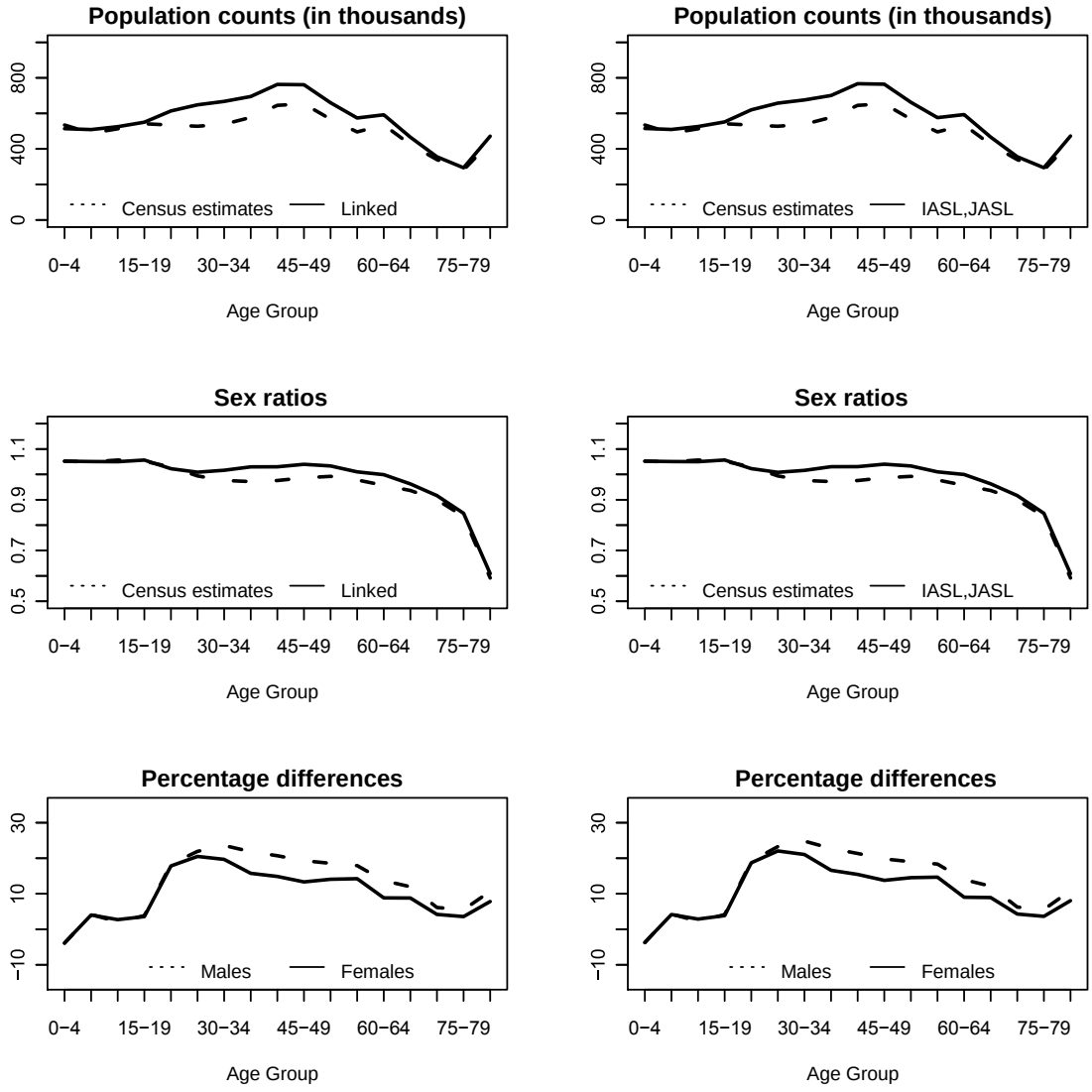


Figure 1: Population counts, sex ratios, and the mean percentage differences between the census estimates by age groups, left panel: the linked data; right panel: the IASL,JASL model

The classic capture-recapture approach (option b) assumes that the probability of inclusion in the Patient Register is independent of the probability of inclusion in the CIS, anyone registered with at least one of the sources is a usual resident, and there are usual residents who are not registered either with the Patient Register or with the CIS. In this case, the total population is equal to

$$N = \widehat{m}_{asl} = \widehat{m}_{++asl} = \widehat{m}_{11asl} + \widehat{m}_{10asl} + \widehat{m}_{01asl} + \widehat{m}_{00asl} \tag{7}$$

where

$$\widehat{m}_{00asl} = \frac{x_{10asl}x_{01asl}}{x_{11asl}}. \tag{8}$$

The saturated capture-recapture log-linear model (IASL,JASL model) for particular age group a, sex s and local authority l produces the same estimated values. For this model the expected population counts $\widehat{m}_{11asl}$, $\widehat{m}_{10asl}$ and $\widehat{m}_{01asl}$ are equal to the observed population counts x_{11asl}, x_{10asl} and x_{01asl} respectively. The equation for the IASL,JASL model is presented in Table 3.

The right panel of Figure 1 presents the census estimates and population counts estimated by the IASL,JASL model, sex ratios of the census estimates and population estimated by the IASL,JASL model, and mean percentage differences between the census estimates and population counts estimated by the IASL,JASL model for males and females by age groups, for the South East region. Like the linked data, the IASL,JASL model also overestimates population counts for almost all of the age groups, and its sex ratio exceeds the census estimates sex ratio for age groups between 25 and 70 year olds.

We continue to investigate the percentage differences between the census estimates; and the linked data and the IASL,JASL model by local authorities. The percentage difference per cell is calculated by $\frac{C_{asl}-\widehat{m}_{asl}}{C_{asl}}$ where C_{asl} and $\widehat{m}_{asl}$ denote the census estimate and the population count estimated by the corresponding model for age a, sex s and local authority l. The percentage differences presented here are grouped according to the local authority quality standards specified in the ONS (2013b) methods and policies report in order to produce comparable maps with the ONS publications. There are seven groups as follows: 'Over 13% and lower', '8.5-13% lower', '3.8-8.5% lower', 'Within 3.8%', '3.8-8.5% higher', '8.5-13% higher', and 'Over 13% higher'. For local authorities coloured in green our estimates are higher than the census estimates, and for local authorities coloured in brown our estimates are lower than the census estimates. Darker colours present higher discrepancies. Finally, the grey local authorities are where our estimates are within 3.8% of the census estimates.

Figure 2 presents the local authority percentage differences for the linked data (on the left panel) and for the IASL,JASL model (on the right panel). The percentage differences are presented for the total population (Figures 2a and 2b), for males (Figures 2c and 2d) and for 20-24 year old males (Figures 2e and 2f). Both the linked data and the IASL,JASL model overestimate most of the local authority census estimates for the selected population groups. The discrepancy is the lowest for total population and it is the highest for 20-24 year old males.

Model	Formula	
I,J,A,S,L	$\log \mu_{ijasl} =$	$\lambda + \lambda_i^I + \lambda_j^J + \lambda_a^A + \lambda_s^S + \lambda_l^L$
IA,IL,JA,JL,AS,AL	$\log \mu_{ijasl} =$	$\lambda + \lambda_i^I + \lambda_j^J + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{ia}^{IA} + \lambda_{il}^{IL} + \lambda_{ja}^{JA} + \lambda_{jl}^{JL} + \lambda_{as}^{AS} + \lambda_{al}^{AL}$
IAL,JAL,AS	$\log \mu_{ijasl} =$	$\lambda + \lambda_i^I + \lambda_j^J + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{ia}^{IA} + \lambda_{il}^{IL} + \lambda_{ja}^{JA} + \lambda_{jl}^{JL} + \lambda_{as}^{AS} + \lambda_{al}^{AL} + \lambda_{ial}^{IAL} + \lambda_{jal}^{JAL}$
IASL,JASL	$\log \mu_{ijasl} =$	$\lambda + \lambda_i^I + \lambda_j^J + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{ij}^{IJ} + \lambda_{ia}^{IA} + \lambda_{is}^{IS} + \lambda_{il}^{IL} + \lambda_{ja}^{JA} + \lambda_{js}^{JS} + \lambda_{jl}^{JL} + \lambda_{as}^{AS} + \lambda_{al}^{AL} + \lambda_{sl}^{SL} + \lambda_{ias}^{IAS} + \lambda_{ial}^{IAL} + \lambda_{isl}^{ISL} + \lambda_{asl}^{ASL} + \lambda_{jas}^{JAS} + \lambda_{jal}^{JAL} + \lambda_{jsl}^{JSL} + \lambda_{iasl}^{IASL} + \lambda_{jasl}^{JASL}$
A_PR	$\log \mu_{++asl} =$	$\lambda + \lambda_a^A + \log(\Gamma_{asl})$ where Γ_{asl} are the three-way table from the Patient Register
A,S_PR	$\log \mu_{++asl} =$	$\lambda + \lambda_a^A + \lambda_s^S + \log(\Gamma_{asl})$ where Γ_{asl} are the three-way table from the Patient Register
AS_PR	$\log \mu_{++asl} =$	$\lambda + \lambda_a^A + \lambda_s^S + \lambda_{as}^{AS} + \log(\Gamma_{asl})$ where Γ_{asl} are the three-way table from the Patient Register
A_IA,IL,JA,JL,AS,AL	$\log \mu_{++asl} =$	$\lambda + \lambda_a^A + \log(\widehat{m}_{++asl})$ where $\widehat{m}_{++asl}$ are the three-way table from the IA,IL,JA,JL,AS,AL model
A,S_IA,IL,JA,JL,AS,AL	$\log \mu_{++asl} =$	$\lambda + \lambda_a^A + \lambda_s^S + \log(\widehat{m}_{++asl})$ where $\widehat{m}_{++asl}$ are the three-way table from the IA,IL,JA,JL,AS,AL model
AS_IA,IL,JA,JL,AS,AL	$\log \mu_{++asl} =$	$\lambda + \lambda_a^A + \lambda_s^S + \lambda_{as}^{AS} + \log(\widehat{m}_{++asl})$ where $\widehat{m}_{++asl}$ are the three-way table from the IA,IL,JA,JL,AS,AL model

Table 3: Log-linear model formulae of the extended capture-recapture models

Figures 1 and 2 show that both the linked data and the saturated model (IASL,JASL model) overestimate the population of the South East region, both at the age group and the local authority levels. The discrepancy between the census estimates and the estimated population is higher for the IASL,JASL model than for the linked data, because the IASL,JASL model adds people to the population captured by the linked data by assuming that there are usual residents who are not registered with either of the registers. To adjust the bias, as a second step, we use log-linear models with offsets to combine population counts of the linked data and the predicted values of the IASL,JASL model with auxiliary information. We modify the AS model presented in Yildiz & Smith (2015) by changing the offset term from direct counts of the Patient Register to the linked data and the predicted values of the IASL,JASL model. The resulting log-linear with offset models combine the three-way tables from the linked data and the IASL,JASL model with the two-way age group-sex marginal table, and the total population count from the auxiliary information.

The log-linear models with offsets are denoted by the symbol of the log-linear model fitted to the auxiliary information and the symbol of the model used as an offset (the linked data or the capture-recapture models) separated by an underscore. The AS model with an offset equal to the linked data is denoted by AS_Linked and has the form

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_s^S + \lambda_{as}^{AS} + \log x_{asl} \tag{9}$$

where

$$x_{asl} = x_{++asl} = x_{11asl} + x_{10asl} + x_{01asl} \tag{10}$$

from the linked data. Similarly, the AS model with an offset equal to the predicted values of the IASL,JASL model is denoted by AS_IASL,JASL and can be written as

$$\log \mu_{asl} = \lambda + \lambda_a^A + \lambda_s^S + \lambda_{as}^{AS} + \log \widehat{m}_{asl} \tag{11}$$

where

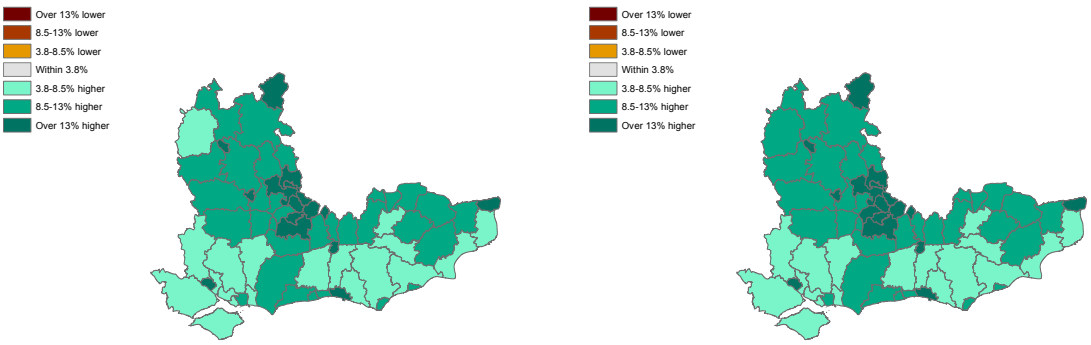
$$m_{asl} = \widehat{m}_{++asl} = \widehat{m}_{11asl} + \widehat{m}_{10asl} + \widehat{m}_{01asl} + \widehat{m}_{00asl} \tag{12}$$

and

$$\widehat{m}_{00asl} = \frac{x_{10asl} x_{01asl}}{x_{11asl}}. \tag{13}$$

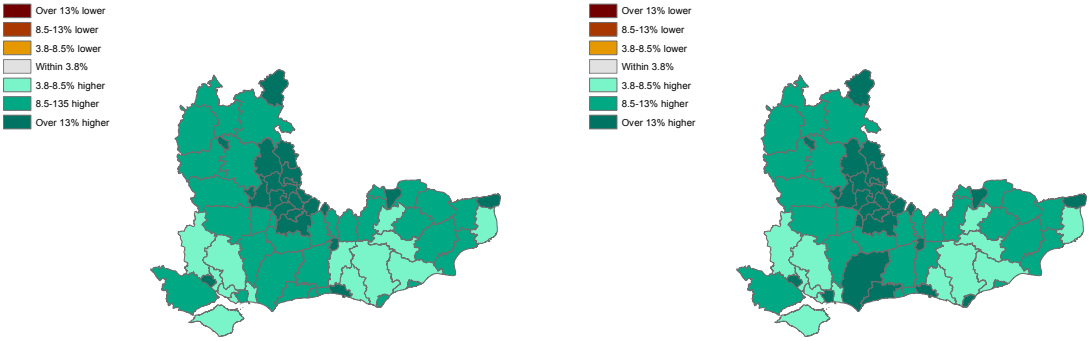
Recall that $\widehat{m}_{ijasl} = x_{ijasl}$ for $(ij) \neq (00)$.

Figure 3a compares the mean percentage differences between the census estimates; and the linked data, the IASL,JASL model, the AS_Linked model, and the AS_IASL,JASL model and shows how much combining the auxiliary age group-sex association structure decreases the discrepancy by age groups. The comparisons are presented for the male population since



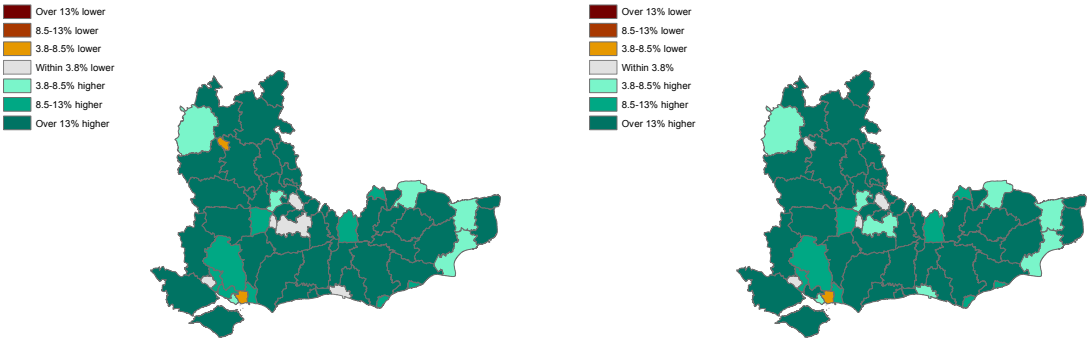
(a)

(b)



(c)

(d)

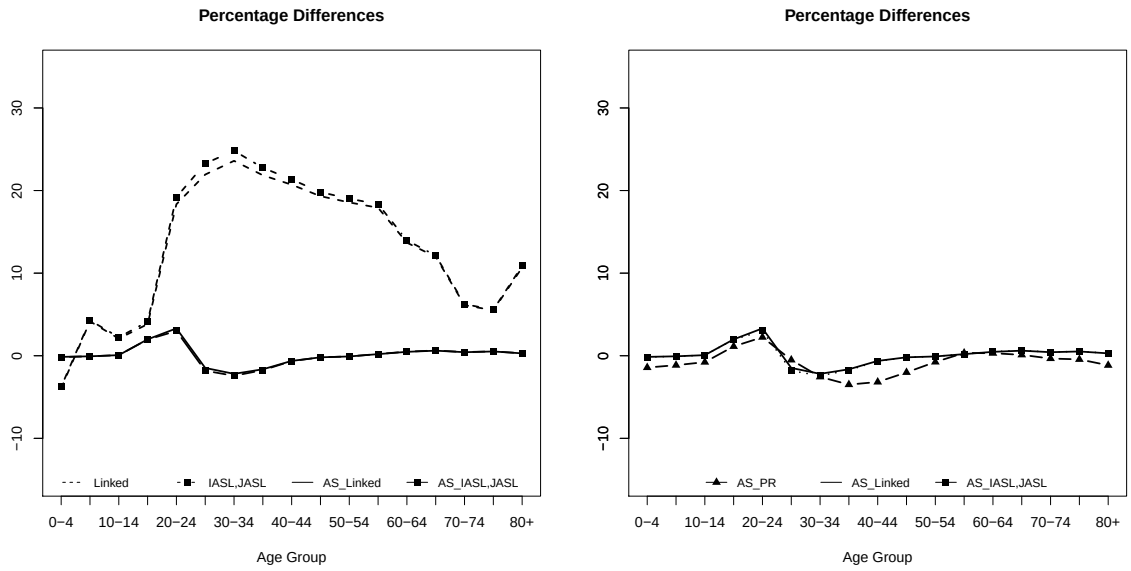


(e)

(f)

Figure 2: Percentage differences between the census estimates and population counts estimated by (a) the linked data and (b) the IASL,JASL model for total population; (c) the linked data and (d) the IASL,JASL model for males; and (e) the linked data and (f) the IASL,JASL model, 20-24 year old males

the discrepancy for males tends to be higher than the discrepancy for females and thus for total population (ONS, 2012, 2013a).



(a) (b)

Figure 3: The mean percentage differences between the census estimates and the linked data (broken line), the IASL,JASL model (broken line with squares), the AS_Linked model (solid line), the AS_IASL,JASL model (solid line with squares) and the AS_PR model (solid line with rectangles) by age groups, males

Figure 3a reveals that including additional information sharply decreases the percentage differences. It can also be seen from the figure that the mean percentage differences for the linked data and the IASL,JASL model are close to each other and the mean percentage differences for the AS_Linked and the AS_IASL,JASL models are close to each other.

The mean percentage differences of the AS_Linked and the AS_IASL,JASL models are also compared with the AS_PR model (Yildiz & Smith, 2015) fitted only to the South East region to show how the discrepancy changes with different offsets in Figure 3b. Both the linked data and the saturated model have less discrepancy than the Patient Register when combined with the auxiliary information (since they provide a better age group-sex-local authority structure than the direct counts from the Patient Register).

Although combining the linked data and the IASL,JASL model with the auxiliary age group-sex structure improves the estimates, these models require the whole five-way table $I \times J \times A \times S \times L$ and assume independent inclusion probabilities. However, the percentage differences with the census estimates change according to age, sex and local authority. Therefore, we continue with different unsaturated capture-recapture log-linear models which allow us to count for dependencies among variables and require less information in the next subsection.

These models have potential to produce more accurate age group sex-local authority association structure than the linked data and the IASL,JASL model. Hence, they may lead to more accurate population estimates when combined with auxiliary marginal tables.

3.2 Parsimonious Capture-Recapture Models

The subsection starts with selecting the appropriate unsaturated / parsimonious capture-recapture models that will explain the higher order structure of the population; and continues with using the estimates from these models as the offsets of the log-linear models to correct the overestimating nature of the capture-recapture models and the linked data.

First, to have a general idea the saturated and the independence (I,J,A,S,L) models are compared. Then, we continue with the all two-way interactions (IJ,IA,IS,AL,JA,JS, JL,AS,AL,SL) model, which is a good starting point for multi-way tables. Later, each pairwise association is removed one by one and the likelihood ratio test was used to decide which pairwise associations are significant and which are not. Since the dataset is large, the likelihood ratio statistic (G^2) of the models are also large. We follow Raymer *et al.* (2009) and use the measure G^2 divided by the residual degrees of freedom (rdf) to control for the relative complexities of the models.

Then, several insignificant pairwise associations are removed to fit simpler models. Afterwards, models with three-way interaction factors are investigated. Table 3 lists the formulae of some of the models, and Figure 4 presents the G^2, residual degrees of freedom (rdf), and $\frac{G^2}{rdf}$ for the models.

Figure 4 shows that out of the two-way interaction models the IA,IL,JA,JL,AS,AL model (in bold) does not have too many two-way interactions, and has relatively lower $\frac{G^2}{rdf}$. The G^2 is also relatively lower than other two-way interaction models which have less pairwise terms. Therefore, we continue our research with this model. However, we also present the mean percentage differences for the IAL,JAL,AS model, (also in bold) to see if including three-way interaction provides better estimates than the IA,IL,JA,JL,AS,AL model for age groups. Among the log-linear models considered the IAL,JAL,AS model has the lowest $\frac{G^2}{rdf}$ value, and it also has the same conditional independence structure as the IA,IL,JA,JL,AS,AL model. According to the IAL,JAL,AS model, the Patient Register and CIS are conditionally independent given local authority and age group; and sex and the variables other than age group are conditionally independent given age group.

Figure 5a shows that the mean percentage differences between the census estimates and the IA,IL,JA,JL,AS,AL model, the IAL,JAL,AS model, and the saturated IASL,JASL model for males are very close to each other. Therefore, we conclude that, there is no need to fit the

Model	G^2	rdf	$\frac{G^2}{\text{rdf}}$
IASL.JASL	0	0	
IAL.JAL.AS	**23,128**	**3,400**	**6.8**
I.J.A.S.L	825,077	6,748	122.3
IA.IS.IL.JA.JS.JL.AS.AL.SL	63,106	5,444	11.6
IA.IL.JA.JS.JL.AS.AL.SL	63,441	5,445	11.7
IA.IL.JA.JL.AS.AL.SL	64,042	5,446	11.8
IA.IL.JA.JS.JL.AS.AL	65,198	5,511	11.8
IA.IS.IL.JA.JL.AS.AL	65,198	5,511	11.8
IA.IL.JA.JL.AS.AL	**65,703**	**5,512**	**11.9**
IA.JA.JL.AS.AL	161,305	5,578	28.9
IA.IL.JA.AS.AL	133,568	5,578	23.9
IA.IL.JA.JL.AS	292,144	6,568	44.5
IA.JA.AS.AL	218,356	5,644	38.7
ISL.JSL.A	630,949	6,416	98.3
IAL.JAL.S	56,267	3,416	16.5
IAS.JAS.L	473,256	6,666	71
IAL.J.AS	209,768	4,538	46.2

Figure 4: Capture-recapture models

saturated model (IASL,JASL) which requires a five-way table. Instead of this, we continue investigating the IAL,JAL,AS and the IA,IL,JA,JL,AS,AL models. Although, fitting the IA,IL,JA,JL,AS,AL and the IAL,JAL,AS models results in close mean percentage differences to the IASL,JASL model for age groups, the discrepancies with the census estimates remain high for all of these three models.

Figure 5b shows that the AS_PR model provides significantly better estimates than the capture-recapture models. Thus, capture-recapture models needs to be modified to provide more accurate population estimates.

Unsaturated/parsimonious capture-recapture models are adjusted by fitting different log-linear models with offsets in the second step. The offsets for these models are the predicted values from the IA,IL,JA,JL,AS,AL and the IAL,JAL,AS models. We evaluate models which use as little auxiliary information as possible since auxiliary three-way marginal tables may not be available in the future. We start with combining the auxiliary age group information (A model) with the selected capture-recapture models, and continue with A,S and AS models.

First, the marginal age group association from the auxiliary data is combined with predicted values provided from the IA,IL,JA,JL,AS,AL model. The corresponding log-linear model is

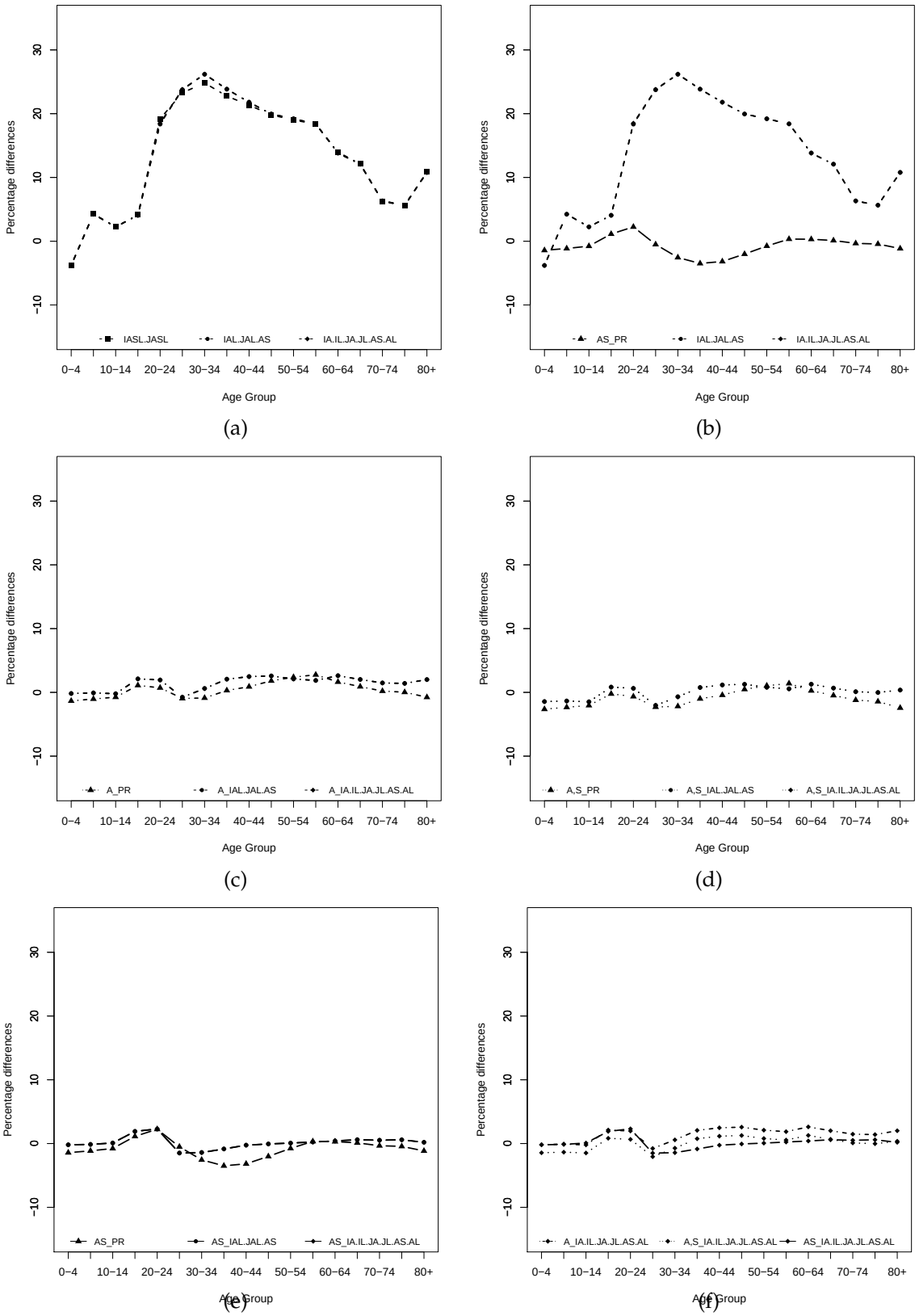


Figure 5: Mean percentage differences between the census estimates and models, males

denoted by A_IA,IL,JA,JL,AS,AL and has the form

$$\log \mu_{asl} = \lambda + \lambda_a^A + \log \widehat{m}_{asl}, \tag{14}$$

where the $\widehat{m}_{asl}$ are predicted values from the IA,IL,JA,JL,AS,AL model marginalized over I and J. Then the A_IAL,JAL,AS model is fitted. The comparison of the mean percentage differences of these models with the A_PR model is presented in Figure 5c. Combining the auxiliary age group information decreases the discrepancy of these models significantly. Although they provide close estimates to the census estimates the mean percentage differences are higher than the A_PR model for the age groups between 15 and 24; between 35 and 49; and for those older than 60 for males.

Next, the A,S_IA,IL,JA,JL,AS,AL model and the A,S_IAL,JAL,AS model are fitted. The comparison of the mean percentage differences of these models with the A,S_PR model is given in Figure 5d. According to Figure 5d, the mean percentage differences between the census estimates and the fitted values of the A,S_IAL,JAL,AS model and the A,S_IA,IL,JA, JL,AS,AL model are lower than the mean percentage differences between the census and the A,S_PR. Thus, when the auxiliary age group and sex association is adjusted, the model fitted to the linked Patient Register and CIS provides better population estimates than the model fitted only to the Patient Register.

Lastly, the AS_IA,IL,JA,JL,AS,AL model and the AS_IAL,JAL,AS model are fitted. The comparison of the mean percentage differences of these models and the AS_PR model is given in Figure 5e. The mean percentage differences between the census estimates, and estimated counts from the AS_IAL,JAL,AS and the AS_IA,IL,JA,JL,AS,AL models are almost equal to zero for most of the age groups; and also lower than the mean percentage differences between the census estimates and the AS_PR for most of the age groups. Thus, we conclude that, when the auxiliary age group-sex association is available it is more sensible to fit a log-linear model with an offset where the offset provides the age group-sex-local authority association from parsimonious capture-recapture model rather than using information only from the Patient Register.

The figures presented above shows that the IA,IL,JA,JL,AS,AL and the IAL,JAL,AS models provide close mean percentage errors for age groups. Therefore, we will compare the models with the auxiliary age group, age group and sex, and age group-sex interaction information with the offsets equal to the fitted values obtained only from the more parsimonious IA,IL,JA,JL,AS,AL model. According to Figure 5f, the AS_IA,IL,JA,JL,AS,AL model provided closer estimates to the census estimates at most of the age groups and the A,S_IA,IL,JA,JL,AS,AL model provided slightly closer estimates to the census estimates at

age groups 15-19 and 20-24.

We continue with evaluating the selected capture-recapture models in terms of percentage differencesby local authorities. Figure 6 compares the percentage differences between the census estimates; and the estimated population counts for the IA,IL,JA, JL,AS,AL model (6a), the A_IA,IL,JA,JL,AS,AL model (6b), the A,S_IA,IL,JA,JL,AS,AL model (6c), and the AS_IA,IL,JA,JL,AS,AL model (6d) for total population. According to Figure 6a, like the linked data and the IASL,JASL model the IA,IL,JA,JL,AS,AL model also overestimates the size of the population. When this model is combined with auxiliary age group information (Figure 6b) the percentage differences for most of the local authorities lie within 3.8% of the census estimates. However, some of the local authorities which were previously slightly overestimated with the IA,IL,JA,JL,AS,AL model (the local authorities which had percentage differences within 3.8 and 8.5%) are now underestimated by the A_IA,IL,JA,JL,AS,AL model. Figure 6c shows that compared to adjusting only age group margins, adjusting both the age group and the sex association does not improve the population count estimates for local authorities. Adjusting the age group-sex association also does not improve the A_IA,IL,JA,JL,AS,AL and the A,S_IA,IL,JA,JL,AS,AL models for local authorities.

Figure 7 presents the percentage differences between the census estimates and the estimated population counts for the IA,IL,JA,JL,AS,AL (7a), the A_IA,IL,JA,JL,AS,AL (7b), the A,S_IA,IL,JA,JL,AS,AL (7c), and the AS_IA,IL,JA,JL, AS,AL (7d) models for the male population. Figure 7a shows that the percentage differences with the census estimates and the population counts estimated by the IA,IL,JA,JL, AS,AL model for males are higher than the percentage differences calculated for the total population (Figure 6a). Adjusting the age distribution decreases the discrepancy in most of the local authorities to within 3.8% level, which means that the age distribution in the linked data is the main cause of the difference between the census estimates and population counts estimated by the IA,IL,JA,JL,AS,AL model for males in local authorities. Adding the sex main effect to the A_IA,IL,JA,JL,AS,AL model by adjusting the age group and sex distribution separately or the age group-sex association does not improve the population count estimates for local authorities. However, from Figure 5f we know that the mean percentage differences for males decrease for most of the age groups when the sex effect is included in the model. The sex distribution in particular local authorities, for which the inclusion of the sex effect causes underestimation, may be different from the general sex distribution in the South East local authorities.

Finally, Figure 8 presents the percentage differences between the census estimates and the estimated population counts for the IA,IL,JA,JL,AS,AL (8a), the A_IA,IL,JA,JL, AS,AL (8b), the A,S_IA,IL,JA,JL,AS,AL (8c), and the AS_IA,IL,JA,JL,AS,AL (8d) models for the male population aged between 20 and 24 years olds. According to Figure 8a, the IA,IL,JA,JL,AS,AL

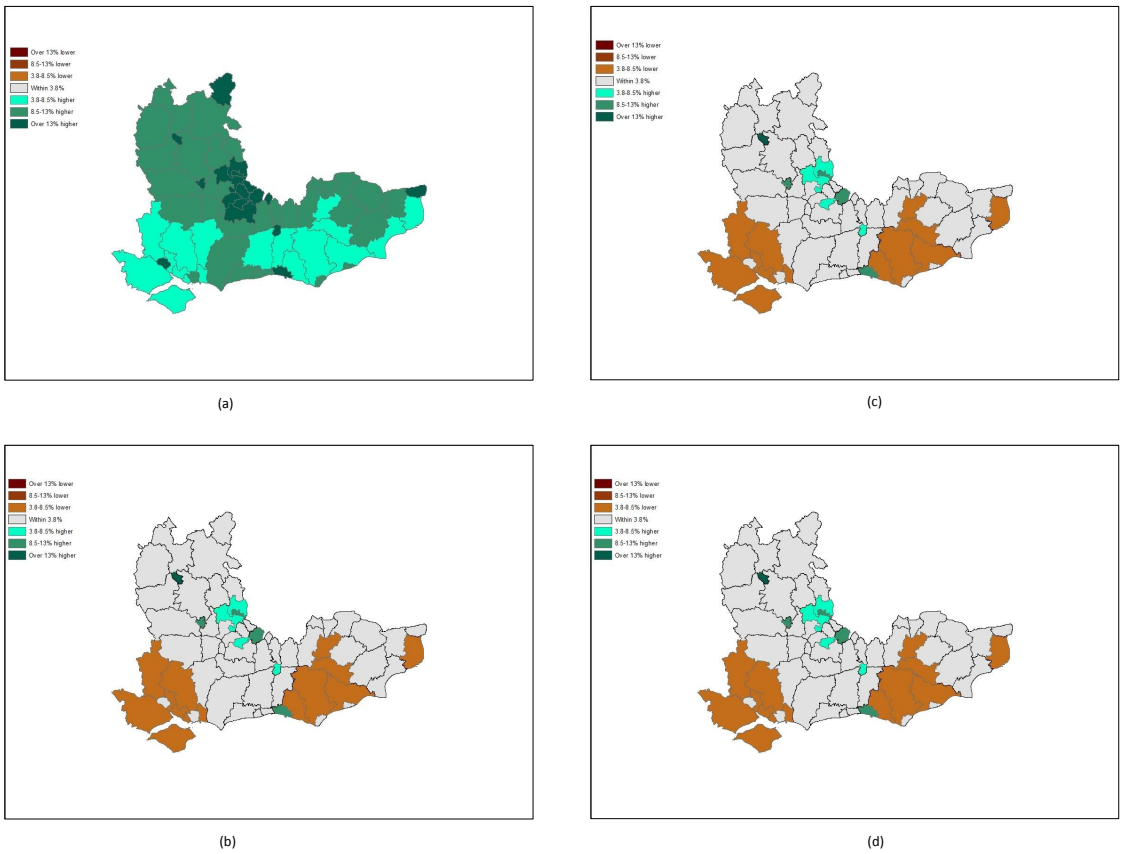


Figure 6: Percentage differences between the census estimates and (a) the IA,IL,JA,JL,AS,AL, (b) the A_IA,IL,JA,JL,AS,AL model, (c) the A,S_IA,IL,JA,JL,AS,AL model, and (d) the AS_IA,IL,JA,JL,AS,AL model, total population

model overestimates the local authority population count for 20-24 year old males by at least 13% for most of the local authorities. This also accords with our earlier observations on the linked data, which overestimated the 20-24 year old male population. Figure 8b also shows that there has been a marked decrease in the discrepancies of the estimated population counts when the predicted values of the IA,IL,JA,JL,AS,AL model are adjusted according to the age group distribution in the South East region. According to Figure 8c, adjusting the sex distribution in addition to age group distribution, results in increasing the number of local authorities with estimated population counts within 3.8% of the census estimates. It is possible that the age group-sex association in the A_IA,IL,JA,JL,AS,AL model was already close to the age group-sex distribution in the census estimates. Therefore, adjusting the age group-sex association with the AS_IA,IL,JA,JL,AS,AL model does not produce better estimated population counts than the A_IA,IL,JA,JL,AS,AL model (8d).

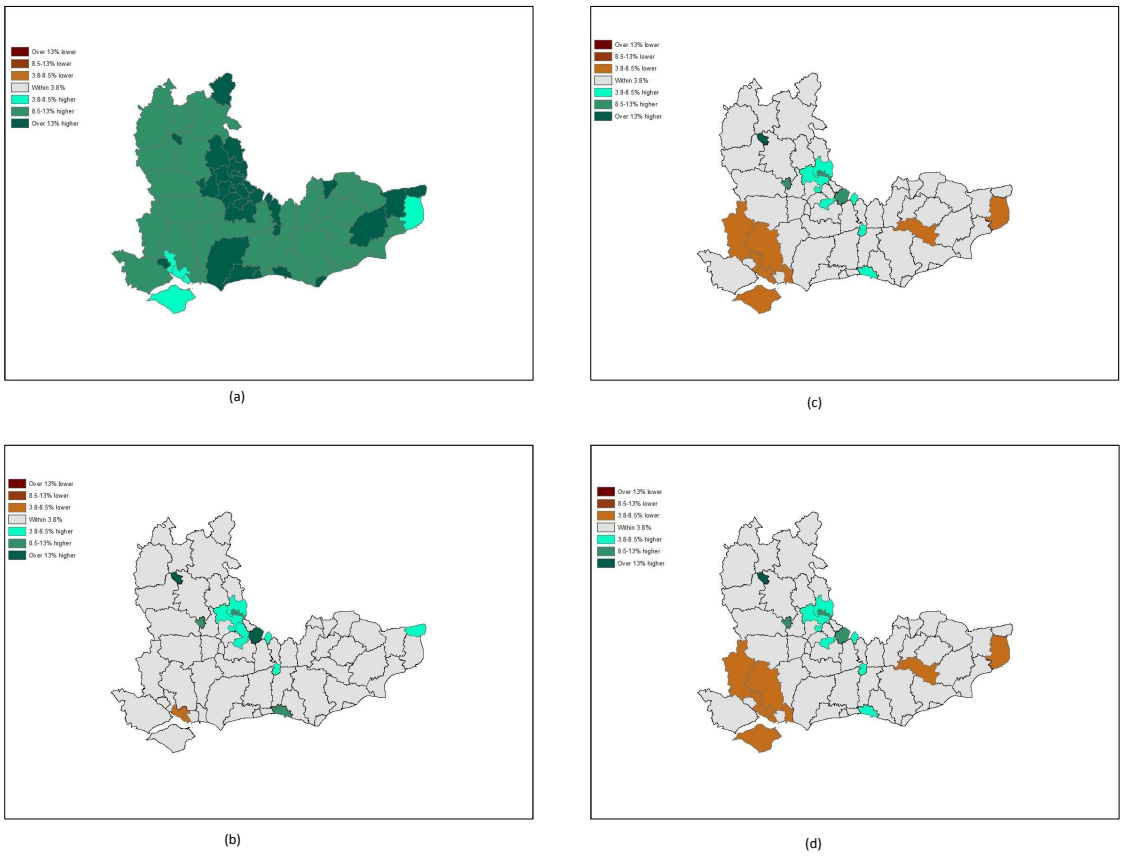


Figure 7: Percentage differences between the census estimates and (a) the IA,IL,JA,JL,AS,AL model, (b) the A_IA,IL,JA,JL,AS,AL model, (c) the A,S_IA,IL,JA,JL,AS,AL model, and (d) the AS_IA,IL,JA,JL,AS,AL model, males

4 Conclusion

This paper presents a model-based approach to estimate population counts using administrative data sources in the absence of both a traditional census and a population register. The methodology makes use of already collected administrative data sources which do not usually aim at collecting information from the usual resident population of a country but a specific population such as pupils in public schools, tax payers or people who registered to vote. These sources are biased, subject to both under coverage and over coverage, and consequently, need to be adjusted when making inferences about the usual resident population.

The proposed methodology has been applied to estimate the usual resident population counts of the South East region of England by age groups, sex and local authority. For this purpose, we used the linked Patient Register and CIS dataset. We demonstrated that the linked data provide a good quality age group, sex, and local authority association structure of the

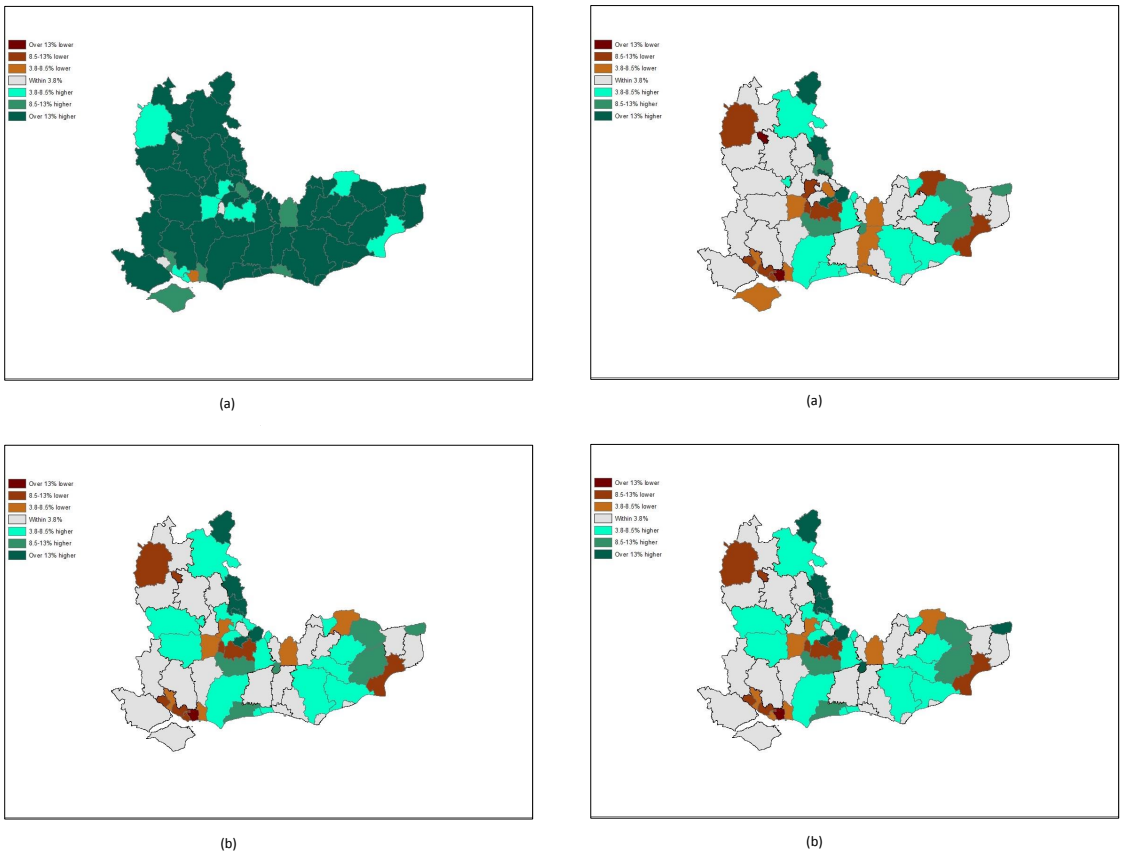


Figure 8: Percentage differences between the census estimates and (a) the IA,IL,JA,JL,AS,AL model, (b) the A_IA,IL,JA,JL,AS,AL model, (c) the A,S_IA,IL,JA,JL,AS,AL model, and (d) the AS_IA,IL,JA,JL,AS,AL model, 20-24 year old males

population than a single source. This is because the linked data not only provide information about people who are registered with at least one of the sources but also allow us to estimate number of people who are not registered with either of the sources by applying the capture-recapture approach. However, when capture-recapture models are used to estimate the size of human populations, the assumptions of the classic capture-recapture approach usually do not hold. Similarly, in this paper, data sources violate both the 'homogeneous inclusion probabilities for at least one source' and the 'no over coverage in sources' assumptions.

The contributions of this paper is threefold. First, we include covariates to capture-recapture models to account for heterogeneity of the inclusion probabilities in the administrative sources. Second, we extend the capture-recapture models by combining the population count estimates of capture-recapture models with auxiliary information to reduce the biased in the final estimates caused by the over coverage in sources. Finally, we improve on Yildiz & Smith (2015) by using different offsets, producing a new set of percentage difference maps comparable to

the ONS publications.

Future research includes, fitting Bayesian log-linear models with offsets and Bayesian capture-recapture models to account for the sampling error when a survey is used as an auxiliary source, and to estimate the uncertainty around the estimates.

References

Agresti, A. 2013. *Categorical Data Analysis*. Third edn. Wiley Series in Probability and Statistics. Wiley.

Bakker, B. F. M., & Daas, P. J. H. 2012. Methodological challenges of register-based research. *Statistica Neerlandica*, **66**(1), 2–7.

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

Brown, J., Abbott, O., & Smith, P.A. 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**(4), 881–906.

Chen, S.X., & Tang, C. 2011. Properties of Census Dual System Population Size Estimators. *International Statistical Review*, **79**(3), 336–361.

Cormack, R. M. 1989. Log-Linear Models for Capture-Recapture. *Biometrics*, **45**(2), 395–413.

Darroch, J. N., Fienberg, S.E., Glonek, G. F., & Junker, B. W. 1993. A three-sample multiple recapture approach to census population estimation with heterogenous catchability. *Journal of the American Statistical Association*, **88**(423), 1137–1148.

Gerritse, S. C, Van Der Heijden, P. G. M., & Bakker, B. F. M. 2015. Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models. *Journal of Official Statistics*, **31**(3), 357–3794.

IWGDMF. 1995. Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development. *American Journal of Epidemiology*, **142**(10), 1047–1058.

ONS. 2009. *Final Population Definitions for the 2011 Census*. Tech. rept. Office for National Statistics.

ONS. 2012. *Beyond 2011: Administrative Data Sources Report: NHS Patient Register*. Tech. rept. Office for National Statistics.

ONS. 2013a. *Beyond 2011: Administrative Data Sources Report: Department for Work and Pensions (DWP) and Her Majesty's Revenue and Customs (HMRC) Benefit and Revenue Information (CIS) and Lifetime Labour Market Database (L2)*. Tech. rept. Office for National Statistics.

ONS. 2013b. *Beyond 2011: Producing Population Estimates Using Administrative Data*. Tech. rept. Office for National Statistics.

Raymer, J., & Rogers, A. 2007. Using age and spatial flow structures in the indirect estimation of migration streams. *Demography*, **44**, 199–223.

Raymer, J., Abel, G., & Smith, P. W. F. 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**(4), 891–908.

Raymer, J., Smith, P. W. F., & Guilietti, C. 2009. Combining census and registration data to analyse ethnic migration patterns in England from 1991 to 2007. *Population, Space and Place*, **17**, 73–88.

Raymer, J., de Beer, J., & Van der Erf, R. 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.

Smith, P. W. F., Raymer, J., & Guilietti, C. 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**(4), 733–753.

Statistics Finland. 2004. *Use of Register and Administrative Data Sources for Statistical Purposes: Best Practices of Statistics Finland*. Tech. rept. Statistics Finland.

Van der Heijden, P. G. M., Whittaker, J., Cruyff, M., Bakker, B., & Van der Vliet, R. 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**(3), 831–852.

Yildiz, D., & Smith, P. W. F. 2015. Models for combining aggregate level administrative data in the absence of a traditional census. *Journal of Official Statistics*, **31**(3), 431–451.

Zhang, L.-C. 2015. On Modelling Register Coverage Errors. *Journal of Official Statistics*, **31**(3), 381–396.

Zwane, E. N., Van der Pal-de Bruin, K., & Van der Heijden, P. G. M. 2004. The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in medicine*, **23**(14), 2267–2281.

VIENNA INSTITUTE OF DEMOGRAPHY

Working Papers

Brzozowska, Zuzanna, Éva Beaujouan and Kryštof Zeman, *Why Has the Share of Two-Child Families Stopped Growing? Trends in Education-Specific Parity Distribution in Low-Fertility Countries*, VID Working Paper 14/2017.

Rengs, Bernhard, Isabella Buber-Ennser, Judith Kohlenberger, Roman Hoffmann, Michael Soder, Marlies Gatterbauer, Kai Themel and Johannes Kopf, *Labour Market Profile, Previous Employment and Economic Integration of Refugees: An Austrian Case Study*, VID Working Paper 13/2017.

Beaujouan, Eva and Caroline Berghammer, *The Gap between Lifetime Fertility Intentions and Completed Fertility in Europe and the United States: A Cohort Approach*, VID Working Paper 12/2017.

Philipov, Dimiter, *Rising Dispersion in Age at First Birth in Europe: Is it related to Fertility Postponement?* VID Working Paper 11/2017 and Human Fertility Database Research Report 2017-005.

Lima, Everton E. C., Kryštof Zeman, Mathias Nathan, Ruben Castro and Tomáš Sobotka, *Twin Peaks: The Emergence of Bimodal Fertility Profiles in Latin America*, VID Working Paper 10/2017 and Human Fertility Database Research Report 2017-004.

Goujon, Anne, Sandra Jurasszovich and Michaela Potancoková, *Religious Denominations in Vienna & Austria: Baseline Study for 2016 - Scenarios until 2046*, VID Working Paper 9/2017.

Winkler-Dworak, Maria, Eva Beaujouan, Paola Di Giulio and Martin Spielauer, *Union Instability and Fertility: A Microsimulation Model for Italy and Great Britain*, VID Working Paper 8/2017.

Testa, Maria Rita and Francesco Rampazzo, *Intentions and Childbearing*, VID Working Paper 7/2017.

Al Zalak, Zakarya and Anne Goujon, *Assessment of the data quality in Demographic and Health Surveys in Egypt*, VID Working Paper 6/2017.

Muttarak, Raya, *Potential Implications of China's 'One Belt, One Road' Strategies on Chinese International Migration*, VID Working Paper 5/2017.

Freitas, Rita and Maria Rita Testa, *Fertility Desires, Intentions and Behaviour: A Comparative Analysis of Their Consistency*, VID Working Paper 4/2017.
