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 $\left\{n_{ij+}\right\}$, $\left\{n_{++k}\right\}$ 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 $\left\{n_{ij+}\right\}$, $\left\{n_{+jk}\right\}$, 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

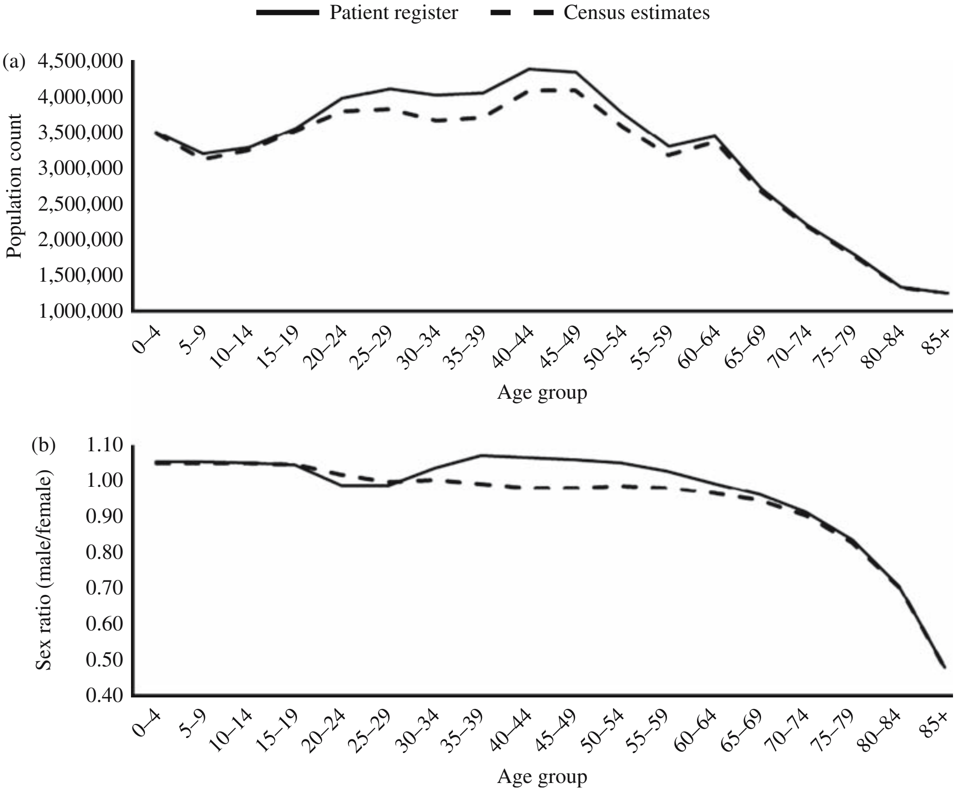

Fig. 1. (a) Population counts by age groups for England and Wales, (b) Sex ratios for the census estimates and the Patient Register (——, Patient Register; – –, Census Estimates)

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	$P_{asr}^{(0)} = \Gamma_{\mathrm{asr}}$		0
Total	Adjusting grand total	$P_{asr}^{(1)} = \Gamma_{\mathrm{asr}} \times \frac{C_{+++}}{\Gamma_{+++}}$	$\log E\left(P_{asr}\right) = \lambda_0 + \log\left(\Gamma_{asr}\right)$	1
AS	Adjusting age group-sex structure	$P_{asr}^{(AS)} = \Gamma_{\mathrm{asr}} \times \frac{C_{as+}}{\Gamma_{as+}}$	$\log E\left(P_{asr}\right) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_{as}^{AS} + \log\left(\Gamma_{asr}\right)$	36
AS,R	Adding region structure to the AS model	$P_{asr}^{(AS,R)^n} = \mathrm{P}_{asr}^{(AS)} \times \frac{C_{++r}}{P_{++r}^{(AS)}}$ $P_{asr}^{(AS,R)^{n+1}} = P_{asr}^{(AS,R)^n} \times \frac{C_{as+}}{P_{as+}}$	$\log E\left(P_{asr}\right) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \log\left(\Gamma_{asr}\right)$	383
AS,SR	Adding sex-region structure to the AS model	$P_{asr}^{(AS,SR)^n} = P_{asr}^{(AS)} \times \frac{C_{+sr}}{P_{+sr}^{(AS)}}$ $P_{asr}^{(AS,SR)^{n+1}} = P_{asr}^{(AS,SR)^n} \times \frac{C_{as+}}{P_{as+}}$	$\log E\left(P_{asr}\right) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \lambda_{sr}^{SR} + \log\left(\Gamma_{asr}\right)$	730
AS,AR	Adding age group-region structure to the AS model	$P_{asr}^{(AS,AR)^n} = P_{asr}^{(AS)} \times \frac{C_{a+r}}{P_{a+r}^{(AS)}}$ $P_{asr}^{(AS,AR)^{n+1}} = P_{asr}^{(AS,AR)^n} \times \frac{C_{as+}}{P_{as+}}$	$\log E\left(P_{asr}\right) = \lambda_0 + \lambda_a^A + \lambda_s^S + \lambda_r^R + \lambda_{as}^{AS} + \lambda_{ar}^{AR} + \log\left(\Gamma_{asr}\right)$	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

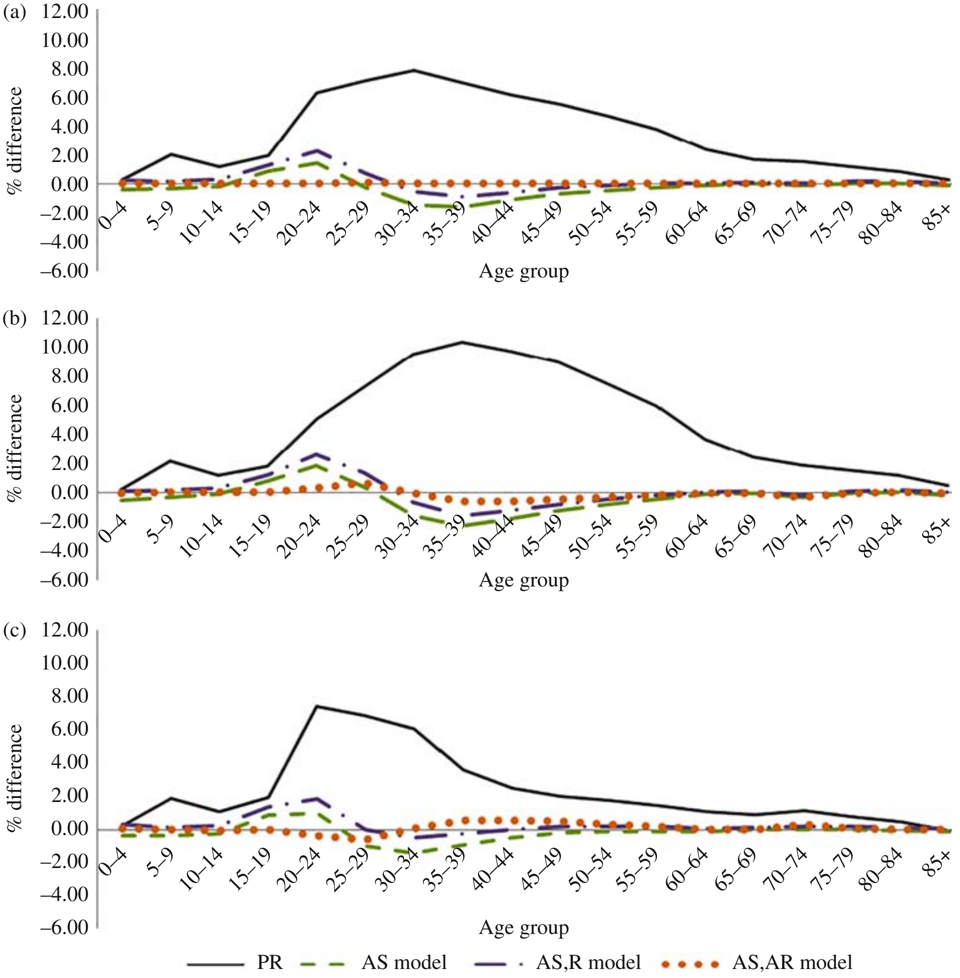

Fig. 2. Mean percentage differences according to age groups for (a) total population, (b) males and (c) females

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

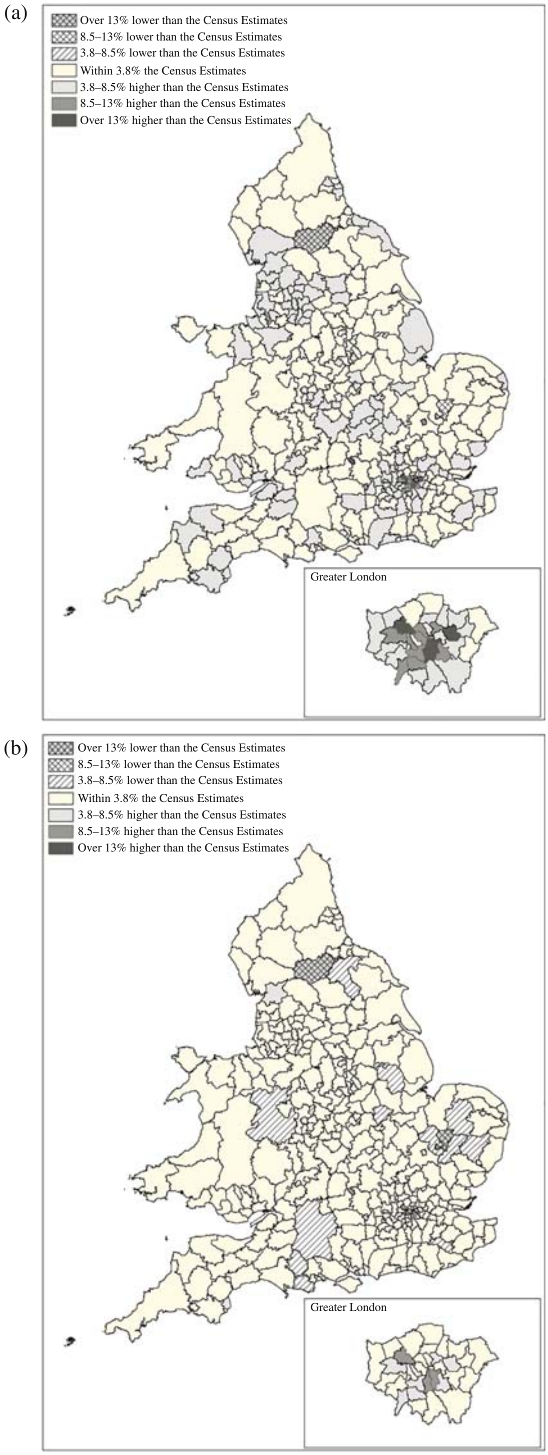
(a)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London
(b)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

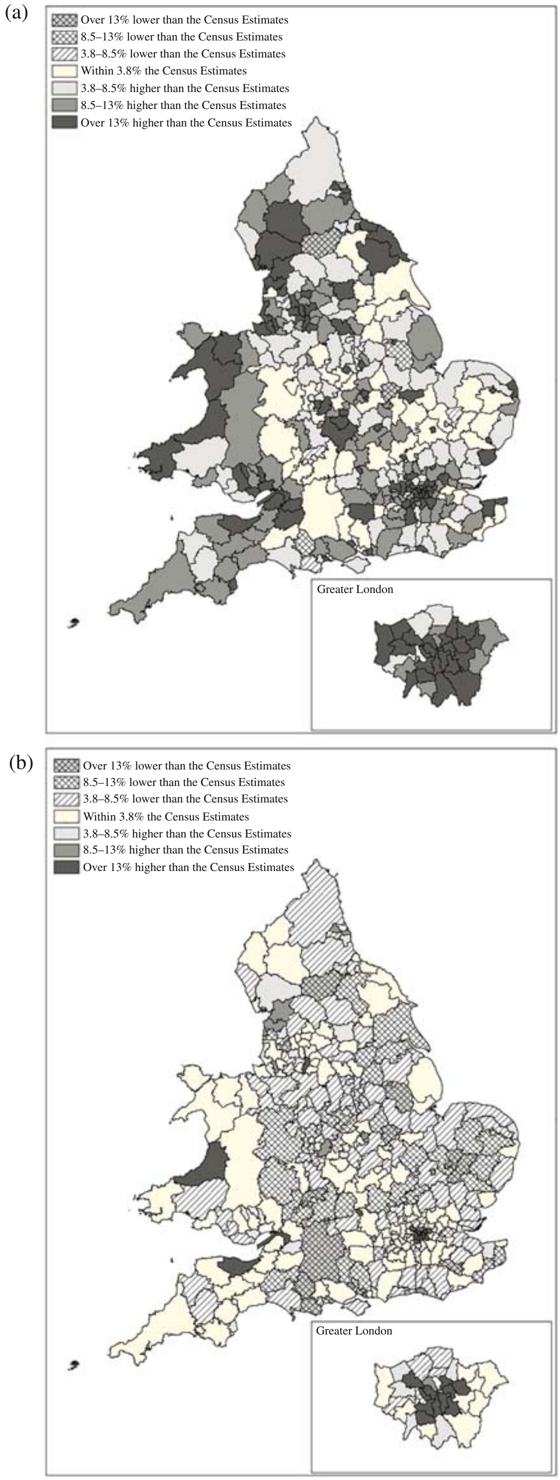
(a)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London
(b)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

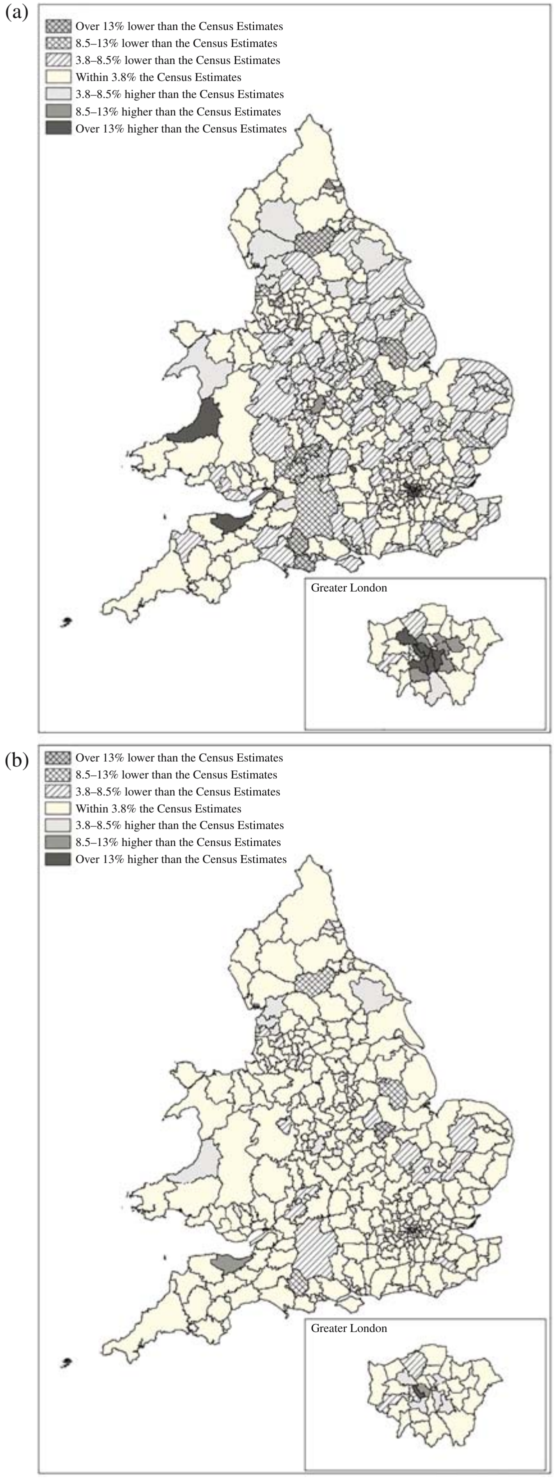
(a)
Over 13% lower than the Census Estimates
8.5–13% lower than the Census Estimates
3.8–8.5% lower than the Census Estimates
Within 3.8% the Census Estimates
3.8–8.5% higher than the Census Estimates
8.5–13% higher than the Census Estimates
Over 13% higher than the Census Estimates
Greater London
(b)
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	+/− 3.8% or better in the peak year
	+/− 6.0% or better in year nine	+/− 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