**Assessing the accuracy of different models for combining aggregate level administrative data**

Dilek Yildiz[[1]](#footnote-1)

Ph.D. student, University of Southampton, UK

Work supervised by: Prof. Peter W.F. Smith1, Prof. Peter van der Heijden1[[2]](#footnote-2)

The Beyond 2011 programme of the Office for National Statistics has been investigating the options either to update the traditional census or to develop alternative methods to produce census statistics. Recently, the National Statistician has recommended “a predominantly online census in 2021 supplemented by the further use of administrative and survey data”. The most comprehensive administrative source in the UK is the National Health Service Patient Register which covers everyone registered with a General Practitioner. However, it is known that the Patient Register is biased and overestimates the total population. The uncertainty in the Patient Register differs for sex, age groups, and local authorities. This paper aims to assess the accuracy of different models for combining the aggregate level 2011 Census estimates of usual residents and 2011 Patient Register. The marginal and the two-way association structures from the census estimates are used to correct the bias in the Patient Register. Population counts by five year age groups, sex and, local authority are estimated by using different log-linear models. The resulting population estimates are compared to the ‘gold standard’ census estimates.

1. **Introduction**

Governments require information about population characteristics in order to make plans, develop policies, and provide public services. At least population counts by age, gender, and some geographical information is required by governments to provide all residents with accessible services such as education and health. One commonly used way of producing the required population information is to take traditional censuses which are valuable sources in terms of producing comprehensive and detailed population information for the whole country. However, at the same time, they are costly and they only provide information in every five or ten years. In addition, the information gathered by a traditional census goes out-of-date in the years following the census. Therefore, the Office for National Statistic (ONS) has been investigating the options either to update the traditional census or to develop alternative methods to produce census statistics to fulfil the needs of the census data users within the Beyond 2011 Programme.

Recently, the ONS [1] reported that “a predominantly online census in 2021 supplemented by the further use of administrative and survey data” has been recommended by the National Statistician and “it is now for the Government and Parliament to determine the arrangements for census-taking in England and Wales in the future”. The UK already has 200 years of experience in traditional censuses and 16% of the respondents in the 2011 Census responded online [1]. Therefore, the main focus is on alternative methods of producing or updating population statistics by making use of already collected administrative data. Unfortunately, the administrative sources assessed by the ONS so far either under count or over count the England and Wales population. Hence, the aim of this research is to assess the accuracy of different models for combining the aggregate level 2011 Census estimates of usual residents and 2011 Patient Register. The marginal and the two-way association structures from the census estimates are used to correct the bias in the Patient Register. Population counts by five year age groups, sex and, local authority are estimated by using different log-linear models. The resulting population estimates are compared to the ‘gold standard’ census estimates.

1. **Data Sources**

This paper deals with one of the most comprehensive administrative data sources in England and Wales, the National Health Service (NHS) Patient Register, which collects information from everyone registered with a General Practitioner. Its high level of coverage and low levels of missing data makes the Patient Register a very valuable administrative source. However, according to the ONS [2] the Patient Register is subject to both under coverage and over coverage, exceeds the census estimates by 4.3% at national level and its sex ratio exceeds the census estimates sex ratio for people aged between 27 and 68. In addition, percentage differences with the census estimates are within 3% for 41% of local authorities. Nevertheless, it is an important data source which has been collecting comprehensive and detailed information from England and Wales population for years. Therefore, if the characteristics of its bias can be well understood and decreased to an acceptable degree, it holds many possibilities to be utilized by the England and Wales statistical authorities in addition to its use in estimating internal migration. One important possibility is using already collected information from the Patient Register to estimate the population of England and Wales either instead of the traditional census or between decennial censuses.

The last census of England and Wales was conducted on 27 March 2011. The 2011 Census counts were adjusted by the ONS for undercount, over count and people counted in the wrong places to produce the 2011 Census estimates by using information collected by the 2011 Census Coverage Survey (CCS) [3]. Therefore, in this paper the census estimates are used instead of the census counts and they are assumed to be the ‘gold standard’.

1. **Methodology**

As mentioned before, the Patient Register is subject both under coverage and over coverage. However, we believe that it holds a good quality age group-sex-local authority association structure which can be corrected easily by combining with marginal tables from a more accurate source. In this research, we used the association structures from the census estimates to correct the Patient Register. Nevertheless, in the absence of a traditional census the association structures will be obtained from alternative sources. One such alternative is the rolling annual surveys as recommended by the ONS [1]. Other possible sources to obtain the association structures include more reliable administrative data sources and other annual surveys. We are interested in estimating the number of people who belong to age group , sex and who are living in local authority . Hence, we used log-linear models (with offsets) to combine the Patient Register and more accurate association structures from the census estimates.

Recently, log-linear models with offsets have been used to combine information from the NHS migration data with the census migration data and the Labour Force Survey to estimate migration flows in [4], [5] and [6]. Raymer et al. combined the NHS migration data with the census migration data to include a variable of interest which is only available in the NHS migration data in [4], and later to model the change in the interregional migration over time in [5]. Again the NHS Patient Register was combined with the 2001 Census and the Labour Force Survey to estimate the migration patterns of the economic activity groups over time in England in [6].

We used unsaturated hierarchical log-linear models with offsets to combine information from the Patient Register and the census estimates to produce accurate and up-to-date population estimates by age group, sex and local authority. We combined two sources by using marginal tables from the census estimates and imposed the three-way association structure from the Patient Register. The marginal tables obtained from the census estimates are used to correct the bias in the Patient Register.

Denote the models by the coefficient of the highest order terms which a variable appears in the model. Assume a three-way table with variables X, Y, and Z, and denote the models by the coefficient of the highest order terms which a variable appears in the model. For example, the mutual independence model is denoted by X,Y,Z and all two way interactions model is denoted by XY,XZ,YZ.

Assume that the three-way table present population counts for 18 five year age groups, two sexes and 348 local authorities in England and Wales and each cell in the table presents the count of people of a certain age group and sex living in a certain local authority. and present the number of people who belong to age group , sex and who are living in local authority in the census estimates and in the Patient Register, respectively. Similarly, denotes estimated population counts for different models, for example is used for the original Patient Register counts and is used for the fitted values of the model denoted by AS.

In this paper we assume that instead of all association structures, only one or two of the age group-sex (henceforth referred as AS), the local authority (referred as L), the sex-local authority (referred as SL) and the age group-local authority (referred as AL) association structures and the population total will be available in the future. Using these association structures from the census estimates, we evaluated the total; the AS; the AS,L; the AS,SL and the AS,AL models with offsets which are presented in Table 1.

Table 1 Log-linear models with offsets

|  |  |  |
| --- | --- | --- |
| Model | Explanation | Log-linear models with offsets |
| PR | The original Patient Register |  |
| Total | Adjusting grand total |  |
| AS | Adjusting age-sex structure |  |
| AS,L | Adjusting local authority structure to the AS model |  |
| AS,SL | Adjusting sex-local authority structure to the AS model |  |
| AS,AL | Adjusting age-local authority structure to the AS model |  |

1. **Comparison of models**

The models are evaluated according to their percentage differences with the census estimates. The percentage differences between the fitted values and the census estimates are calculated by.

Figure 1 shows mean percentage differences for age groups for total population, males and females. Mean percentage differences for age groups for the total model follow the same pattern as the original Patient Register percentage differences but at a lower level (not shown here). The same pattern can also been seen in mean percentage differences for both males and females separately. The AS model decreases absolute percentage differences for all age groups except the youngest age group which was already very accurate in the Patient Register. Mean percentage differences for the AS,L and the AS,SL models are very close to each other both for total population and for two sexes therefore the AS,SL models are not plotted. The AS,AL model provides almost a perfect fit for the total population with the highest percentage difference 0.06 for the 25-29 year old age group. For males AS,AL model overestimates the 25-29 age group and underestimates 35-39, 40-44 and 45-49 age groups slightly. Unsurprisingly, by constraining the population total at constant results in the underestimation of 25-29 age group and overestimation of 35-39, 40-44 and 45-49 age groups, for females.

Table 2 presents the percentage of local authorities within 3.8 % of the census estimates for selected age group-sex population groups. According to Table 2, for the AS,AL model more than 75% of all local authorities are within 3.8% of census estimates for the population groups considered in this paper. Moreover, the AS,AL model results in more than 85% of all local authorities within 3.8% of census estimates except for the 20-24 year olds.

Table 2 Percentage of local authorities within 3.8 percent of the census estimates

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PR | Total | AS | AS,L | AS,SL | AS,AL |
| 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 |

(a)

(b)

(c)

Figure 1 Mean percentage differences according to age groups for (a) total population, (b) males and (c) females

1. **Conclusion**

In this paper we tried to correct the bias in the Patient Register by using the association structures from the 2011 Census Estimates. However, the population counts from the census estimates may not be available in the future. In this case, it is possible to obtain association structures from another source such as another administrative source or annual surveys. Even if decennial censuses continue, the methodology proposed here will allow the estimation of population counts for the years between two censuses.

According to our research the most efficient model to decrease the discrepancy between the Patient Register and the census estimates is the AS,AL model. It improves the Patient Register in terms of percentage differences at local authority and age group level.

As a conclusion this research evaluates different log-linear models in terms of their capacity of producing accurate population counts for age group, sex and local authority groups. On the other hand, it also presents a perspective to understand which population groups are tend to be missed by the Patient Register and how much the Patient Register can be improved just by combining it with the specific association structures. This research can be extended in a way that known issues about particular administrative sources and expert knowledge can be taken into account to develop different models. Employing the best resulted models for different age groups and sex, and combining the resulting estimates to produce accurate population counts is also possible.

1. **References**

[1] Office for National Statistics (ONS) (2014), 27 March 2014 - The census and future provision of population statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority, Office for National Statistics.

[2] Office for National Statistics (ONS) (2012), Beyond 2011: Administrative Data Sources Report: NHS Patient Register, Office for National Statistics.

[3] Office for National Statistics (ONS) (2012), The 2011 Census coverage assessment and adjustment process, Office for National Statistics.

[4] Raymer, J., Abel, G. and 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, 170(4), 891-908.

[5] Raymer, J., Smith, P.W. F., and 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.

[6] Smith, P.W. F., Raymer, J., and Guilietti, C. (2010), Combining available migration data in England to study economic activity flows over time, Journal of the Royal Statistical Society, Series A, 173(4), 733-753.

1. Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom [↑](#footnote-ref-1)
2. Utrecht University, The Netherlands

   Acknowledgment: This research is funded by the joint Economic and Social Research Council - Office for National Statistics studentship. [↑](#footnote-ref-2)