**Towards an integrated census-administrative data approach to item-level imputation for the 2021 UK Census**

Fern Leather, Katie Sharp, Steven Rogers1

**Abstract**

In preparation for 2021 UK Census the ONS has committed to an extensive research programme exploring how linked administrative data can be used to support conventional statistical processes. Item-level edit and imputation (E&I) will play an important role in adjusting the 2021 Census database. However, uncertainty associated with the accuracy and quality of available administrative data renders the efficacy of an integrated census-administrative data approach to E&I unclear. Current constraints that dictate an anonymised ‘hash-key’ approach to record linkage to ensure confidentiality add to that uncertainty. Here, we provide preliminary results from a simulation study comparing the predictive and distributional accuracy of the conventional E&I strategy implemented in CANCEIS for the 2011 UK Census to that of an integrated approach using synthetic administrative data with systematically increasing error as auxiliary information. In this initial phase of research we focus on imputing single year of age. The aim of the study is to gain insight into the point at which the degree of error in the administrative data leads to poorer performance than a conventional E&I strategy such as that used for 2011 UK Census.

Key words: census, imputation, administrative data, canceis

1. **Introduction**

Adjusting census data for item-level inconsistencies and non-response (from here on referred to as imputation) is an important part of census data processing. The 2011 UK Census imputation strategy was a donor-based approach implemented in CANCEIS (Aldrich and Rogers, 2012; Wardman and Rogers, 2012; Wardman et al., 2014).

In CANCEIS, the variables within a module, along with other auxiliary information, served as matching variables representing the underlying imputation model. A Nearest Neighbour algorithm constructed a pool of potential donors statistically similar to the record being imputed based on available data in the matching variable set. From that pool, the donor that supplied an imputed value was chosen randomly from a set of near minimum change imputation actions (NMCIAs) (Bankier et al., 1999; CANCEIS, 2009). This strategy provides point and variance estimates of the distribution of the missing data conditioned by the underlying imputation model (from here on referred to as imputation estimates).

Following a detailed research programme after completion of the 2011 Census, the National Statistician recommended a predominantly online UK Census in 2021, supplemented by the use of administrative data. This was approved by the UK Government in July 2014. The ONS is now fully committed to the development of a different kind of Census in 2021. Exploring the impact that administrative data may have on the accuracy of imputation estimates is therefore an important facet of that commitment.

Blum (2006) describes a number of mechanisms by which administrative records can aid the imputation process. These include cold-deck imputation, improving model specification, and continuous quality assurance. The first of these, cold-deck imputation, carries a risk that observed census data could be changed on the basis of inconsistent values obtained from an external administrative source. Given this risk and the uncertainty surrounding both the quality of available administrative sources and the linkage mechanism, our initial research has focused on the use of linked administrative data as auxiliary information within the conventional hot-deck approach already employed by ONS for imputation of census data.

The aim of the research was to determine whether linked administrative data could improve the accuracy of imputation estimates for age, and to gain some insight into the threshold at which the degree of error in the administrative data leads to poorer performance than a conventional E&I strategy, with a view to eventually establishing some general principles for the use and evaluation of administrative sources in the 2021 UK Census.

Here we present findings of some preliminary work based on 2011 UK Census data, aimed at evaluating the differences between imputation estimates obtained by using a donor based imputation strategy that includes synthetic administrative data as auxiliary information in the form of an ‘admin age’ variable linked to each census record, compared with the conventional approach as used for the 2011 UK Census.

The analysis focuses on assessing the predictive and distributional accuracy (Chambers, 2001) of the imputation estimates for one ‘typical’ Census delivery group that had age randomly perturbed for 5% of records and was imputed under the following conditions: no administrative data (2011 Census approach), exact administrative data, administrative data with +/-3 years error, administrative data with +/-6 years error, administrative data with +/-12 years error, and finally, administrative data using observed error distributions from two real administrative datasets.

Demographic data for households containing 1 to 6 people were used, with the age variable imputed using the 2011 UK Census imputation model implemented in CANCEIS, which included the matching variables of: relationship to household person 1; sex; marital status; activity last week; student indicator, term-time address indicator; country of birth indicator; ‘hard to count’ rating; as well as the additional matching variable of ‘admin age’ provided by the synthetic administrative sources, with its weight set to equal that used for age in the 2011 Census. Admin age was set to equal the true census age for all clean records.

The final stage of the study involved deriving error distributions for age from real administrative datasets by matching them to 2011 Census data (assumed to represent the true age) by forename, surname, sex and postal code. These error distributions were then used to create additional synthetic admin age variables which were used to impute the perturbed dataset, as above, in order to determine where real administrative datasets might fall on the continuum of error.

1. **Results & Discussion**
	1. **Predictive accuracy**

Figures 2.1-1 to 2.1-5 illustrate the distributions of the error in the imputed age compared with the true age for each of the experimental conditions, based on all potential donors. The exact administrative data condition clearly performed better than the no administrative data condition, with errors centred around zero but with a much lower standard deviation (3.17, compared with 11.89 for the no administrative data condition, see Table 2.1-1), while the other administrative data conditions produced non-normal distributions of error, with modes reflecting the level of error in the admin data variable.

Figure 2.1-1. Figure 2.1-2.

Exact administrative data condition No administrative data condition

 

**Figure 2.1-3. Figure 2.1-4.**

**+/- 3 years administrative data error condition +/- 6 years administrative data error condition**

** **

**Figure 2.1-5.**

**+/- 12 years administrative data error condition**



Table 2.1-1.

Summary statistics for the error in the imputed age compared with the true age for the no admin data and exact admin data conditions

|  |  |  |
| --- | --- | --- |
|   | **No admin data** | **Exact admin data** |
|   | ***Estimate*** | ***95% Confidence Limits*** | ***Estimate*** | ***95% Confidence Limits*** |
| Mean | 0.14 | 0.10 | 0.19 | 0.21 | 0.18 | 0.23 |
| Std Deviation | 11.89 | - | - | 3.17 | - | - |
| Variance | 141.38 | - | - | 10.05 | - | - |

* 1. **Distributional accuracy**

Recovery of the true age distribution of the perturbed data was also assessed. Table 2.2-1 illustrates summary statistics for the error in the counts by single year of age for each condition.

**Table 2.2-1.**

**Measures of distributional accuracy for each experimental condition**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Error | Exact admin data | No admin data | Admin +/-3 | Admin +/-6 | Admin +/-12 |
| Mean | -0.019 | -0.028 | -0.086 | -0.114 | -0.065 |
| Mean 95% CI | -4.40 , 4.36 | -4.90 , 4.84 | -6.01 , 5.84 | -7.99 , 7.76 | -6.98 , 6.85 |
| St Dev | 22.43 | 25.40 | 30.61 | 40.71 | 36.26 |
| SSE | 51,294 | 68,387 | 97,425 | 172,354 | 140,709 |
| RMSE | 22.32 | 25.28 | 30.46 | 40.52 | 36.10 |
| RMSE 95% CI | 18.57 , 25.52 | 21.04 , 28.91 | 22.99 , 36.43 | 30.31 , 48.62 | 30.57 , 40.88 |

The exact administrative data condition performed better than the no administrative data condition, with a lower mean error by age, lower sum of squared errors and lower root mean squared error, indicating both less absolute error over the whole age distribution and less error for each individual age. The other administrative data conditions on the other hand, had particular difficulty in recovering the true age distribution at the school age/working age boundary between 15 and 16 years old. As table 2.2-2 shows, only 60% of the true number of 16 year olds were imputed for the +/- 6 years error condition, whereas both the no administrative and exact administrative data conditions accurately recovered the true count of 16 year olds.

**Table 2.2-2**

**Post-imputation frequencies of 16 year olds for each experimental condition**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|   | **True values** |  | **Exact admin data** |  | **No admin data** |  | **Admin +/-3** |  | **Admin +/-6** |  | **Admin +/-12** |
| **Age** | **n** | **% of total** |  | **n** | **% of total** |  | **n** | **% of total** |  | **n** | **% of total** |  | **n** | **% of total** |  | **n** | **% of total** |
| 16 | 368 | 1.2 |   | 372 | 1.2 |   | 375 | 1.2 |   | 267 | 0.9 |   | 215 | 0.7 |   | 348 | 1.1 |

This appears to be due to the presence records with a true age on one side of the school age/working age boundary and an admin age on the other side. The systematic differences between school age and working age records and the CANCEIS parameterisation used in the 2011 Census to maintain these differences meant that, for example, it was difficult for records with a true perturbed age of 10 and an admin age of 16 to match to 16 year old potential donors, but for records with a true age of 16 and an admin age of 22, there was little to prevent a match to 22 year old donors, resulting in a net decrease in 16 year olds compared with the true distribution.

The +/- 12 years error condition performed better than the +/-6 years condition in terms of distributional accuracy, both in terms of absolute error over the whole distribution and at the school age/working age boundary. This appears to be because the error in the admin age variable was so great that it could not be matched on closely while maintaining consistency with the observed data, resulting in consistency being prioritised over matching on admin age, meaning that there was a higher chance of recovering the true age distribution. In this situation, therefore, the observed Census data is effectively protected from high levels of error in the administrative data.

* 1. **Other measures**

Some other key measures of imputation accuracy were also analysed (Table 2.3-1). Administrative data of any quality led to a reduction in the size of the donor pool and restricted the age range of the potential donors to only those that matched the admin age very closely. The +/- 12 years condition restricted the donors less than the other conditions for the reasons stated in section 2.2 – at this level of error it was simply not possible to find donors that matched closely on admin age while maintaining consistency with the observed variables for a large number of records.

As would be expected, the percentage of records that would fail the edit checks if the admin age was directly substituted into the census age increased as the administrative error increased, reaching almost 20% for the +/- 12 years condition. This is important because the observed census data are assumed to be the ‘truth’ so, if the administrative data are inconsistent with them, there is a risk that observed data could be changed on the basis of the administrative data if the conventional edit rules are implemented. The fact that the imputation method implemented in CANCEIS is able to protect observed data from inconsistent administrative data is therefore an advantage over a direct substitution method.

**Table 2.3-1**

**Other measures of imputation accuracy**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | **Exact admin data** | **Admin +/-3** | **Admin +/-6** | **Admin +/-12** | **No admin data**  |
| Total size of donor pool  | 56,077 | 57,529 | 59,807 | 70,068 | 258,116 |
| Average age range of potential donors (years)  | 1.0 | 1.1 | 1.4 | 2.5 | 14.3 |
| Percentage of records where true age is within range of imputation actions  | 53.1 | 20.9 | 12.9 | 20.5 | 58.3 |
| Percentage of records failing edit checks if admin age directly substituted into census age  | 0 | 3.8 | 9.1 | 19.4 | N/A |

* 1. **Error distributions derived from actual administrative datasets**

This initial phase of research also involved analysing the error distributions of two 2011 extracts of actual administrative datasets, the NHS Patient Register (PR) and the Department for Work and Pensions’ Customer Information System (CIS), linked to 2011 Census data by forename, surname, sex and postal code. Where age was observed on the administrative source and the Census, both the PR and CIS datasets had over 98% exact agreement with the Census age, indicating high reliability of the age variable.

These observed error distributions were then used to construct a further set of synthetic administrative datasets, which were used as auxiliary information in imputation of the perturbed dataset, in the same way as with the other synthetic datasets.

Table 2.4-1 sets out the distributions of the error in the imputed age compared with the true age for the two conditions, compared with the exact administrative data and no administrative data conditions. Predictive accuracy was similar to that of the exact administrative conditions, which is to be expected given the high reliability of the age variable.

**Table 2.4-1**

Summary statistics for the error in the imputed age compared with the true age for the CIS and PR synthetic conditions contrasted with the no admin data and exact admin data conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | PR | CIS | No admin data | Exact admin data |
|   | ***Estimate*** | ***95% Confidence Limits*** | ***Estimate*** | ***95% Confidence Limits*** | ***Estimate*** | ***95% Confidence Limits*** | ***Estimate*** | ***95% Confidence Limits*** |
| Mean | 0.20 | 0.17 | 0.23 | 0.21 | 0.18 | 0.24 | 0.14 | 0.10 | 0.19 | 0.21 | 0.18 | 0.23 |
| Std Dev | 3.32 | - | - | 3.33 | - | - | 11.89 | - | - | 3.17 | - | - |
| Variance | 11.02 | - | - | 11.10 | - | - | 141.38 | - | - | 10.05 | - | - |

The distributional accuracy for both conditions was also higher than the no administrative condition, with RMSE figures of 21.7 and 21.4 for the PR and CIS conditions respectively, compared with 25.3 for the no administrative condition.

The results therefore show that it is possible for administrative data to improve the accuracy of imputation estimates for age under certain conditions and that readily available administrative datasets appear to be of sufficiently high quality to be able to deliver this improvement. The nearest neighbour minimum change methodology implemented in CANCEIS also ensured that, where the administrative data was so erroneous as to be inconsistent with the observed data, consistency was maintained over matching closely on the admin age variable. This effectively protected the observed census data from high levels of error in the administrative data, demonstrating a clear advantage over alternative direct substitution methods.

Going forward, the insight gained from this research can be used to evaluate other potential administrative datasets and other variables that had higher imputation rates in the 2011 Census but may be less reliable in administrative sources. Preliminary research carried out by other NSIs suggests that administrative records may be a promising avenue for other more difficult variables, like race (Rastogi et al., 2014).

The study assumed perfect linkage and 100% coverage in order to focus on the effects of error in the administrative data, but this would not be achieved in a real situation so the findings represent a ‘best-case’ scenario. Further work will therefore be required to determine the effects of imperfect linkage and less than 100% administrative data coverage to allow a full cost-benefit analysis of the method to be carried out in preparation for the 2021 Census.

**References**

Aldrich, S., Wardman, L., and Rogers, S. (2012), The practical implementation of the 2011 UK Census imputation methodology. Available online: [06/01/2015] <http://www.unece.org/stats/documents/2012.09.sde.html#/>

Bankier, M., Lachance, M., and Poirer, P. (1999), A generic implementation of the new imputation methodology. Accessed online: [12/01/2016] <http://www.ssc.ca/survey/documents/SSC2000_M_Bankier.pdf>

Blum, O. (2006), Evaluation of editing and imputation supported by administrative records. Statistical Data Editing Volume 3: Impact on Data Quality, UNECE, pp300-309. Available online: [10/12/2015] <http://www.unece.org/stats/publications/editing/SDE3.html>

CANCEIS (2009), Users Guide V4.5. CANCEIS Development Team. Social Survey Methods Division, Statistics Canada

Chambers (2001), National Statistics Methodological Series Report 28: Evaluation criteria for statistical editing and imputation, Office for National Statistics

Rastogi, S., Fernandez, L., Noon, J., Zapata, E. and Bhaskar, R (2014). Exploring administrative records use for race and hispanic origin item non-response. US Census Bureau CARRA Working Paper Series #2014-16. Available online: [01/04/2016] <https://www.census.gov/srd/carra/16_Exploring_Admin_Recs_for_Race_and_Hispanic_Origin_Item_NR.pdf>

Wardman, L., Aldrich, S., and Rogers, S. (2012), Item imputation of Census data in an automated production environment; advantages, disadvantages and diagnostics. Available online: [06/01/2015] <http://www.unece.org/stats/documents/2012.09.sde.html#/>

Wardman, L., Aldrich, S., and Rogers, S. (2014), 2011 Census item edit and imputation process. Available online: [06/01/2015] <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-and-imputation-process.pdf>