Coverage Adjustment methodology for the 2011 Census

James Brown (UoS)[1]
Christine Sexton, Alan Taylor and Owen Abbott (ONS)

1) Introduction

The one-number census imputation system developed in Steele, Brown and Chambers (2002) was able to create an adjusted database in 2001 but it has been recognised that this was not without some practical difficulty in achieving constraints. However, as with other parts of the coverage assessment process and planning for 2011, the starting point is the working system from 2001. Previously (see Abbott and Brown, 2007) we outlined the ambitious development of a new system for this most specialised component of the whole process. However, the constraints of resource have meant that this alternative approach could not be pursued. Therefore, in this paper we return to reviewing the 2001 system with proposed incremental improvements that fit more naturally within the 2001 framework, and that will allow much of the system to be re-used or developed rather than requiring major new software components to be developed to implement a whole new approach.

In the following section we give a brief review of the 2001 approach followed by section 3 that presents a series of incremental improvements we wish to evaluate as improvements to the system, building on the position that the 2001 system is usable but perhaps not ideal. Results from studies we have undertaken to evaluate the 2001 performance in comparison with proposed improvements are outlined in section 4 while in section 5 we outline the outstanding issues.

2) Review of 2001 System

At the first stage we modelled missed households (including the individuals within them) and then missed individuals within counted households at the estimation area (EA) level. The EAs were groups of contiguous Local Authorities (LAs) with populations of around half a million and the CCS was designed to give high quality estimates of the population by age

[1] James Brown's involvement was part funded via ADMIN at the Institute of Education, which is funded by the ESRC under their NCRM Methods Node programme (ref: RES-576-25-0014).

and sex at this level. The two coverage models for households and individuals within counted households were multinomial logistic regressions that allowed for the households (or individuals within counted households) to be counted by the census and the CCS or to just be counted by one or other. From these two models we derived the census coverage probabilities for households (by characteristics) and individuals within counted households (by characteristics). As these coverage probabilities varied by several characteristics but did not allow for those households or individuals missed by both, they were calibrated to marginal estimates at the EA level for households and at the LA level (by age and sex) for persons. These marginal estimates used a dual-system methodology so the calibration exercise was not only for consistency but also ensured the coverage probabilities reflected households and individuals missed by both the Census and CCS.

At the second stage we imputed the completely missed households including all the individuals contained within them. This was based on some basic household characteristics such as tenure as well as the characteristics of the individuals within the households through a simplified household structure variable. Placement of households was partly driven by locations of dummy forms within LAs or was random when there were no appropriate dummy forms. At the third stage we imputed missed individuals into counted households. This required us to find donor individuals and appropriate households to place them in. The fact that individuals entered into the database at two points required a fourth stage that adjusted the age-sex by household size distribution to ensure consistency between the final database and agreed estimates at the LA level. In the following we discuss some issues that were encountered when the system was implemented. These are discussed in more detail in Abbott and Brown (2002).

2.1) 'Problem One' – The missed household model did not control the size of households very efficiently (or the other characteristics of the imputed individuals).

The multinomial model for missed households had a simplified variable for household structure that did not contain lots of detail with respect to the number of household members or their detailed age-sex structure, let alone control on variables like 'activity last week'. A direct consequence of that in 2001 was the household imputation adding too many individuals within certain age-sex groups before we had even done the second imputation stage. There is a trade-off here between geographic variation and characteristics. In 2001 the choice was to model at relatively low levels of geography (independent models within each EA) and therefore the models were relatively simple with respect to household types and size. Further analysis of the 2001 data suggests that it would be better to fit more detailed models at say a regional level and then constrain the weights to marginal totals at the LAD/EA level. This would allow more detailed definitions of variables within the models but would still reflect differences across geography through the calibration and the inclusion of LA effects in the models.

Gaining more control over the individuals imputed at the household stage is crucial to remove (or reduce) the need for the fourth stage from 2001 as this is the most computer intensive stage and often involved forcing changes to the database to meet our key totals.

2.2) 'Problem Two' – How do we ensure the final database is consistent with census edit rules?

The approach in 2001 achieved this by copying records that had already satisfied the census edit rules. The advantage of this was that the editing and item non-response could be run in parallel with the matching and estimation. (A final run was required for relationships within households that had gained imputed members.) However, creating exact copies onto the database does not necessarily create a database that really reflects the heterogeneity in the population. In addition, the imputation models only controlled a restricted set of characteristics (age, sex, tenure, etc) and the system relied on interrelationships between variables within records to control other characteristics. This is essentially what the item non-response system does but we would expect it to do it more efficiently. Therefore, if there is sufficient time and resource, it seems sensible to use the models to create the basic household and individual characteristics and fill in the detail using the standard item non-response imputation system. Not only will this preserve interrelationships between variables but it would avoid the simple copying of entire records.

3) Developing the System

In this section we present the incremental improvements to the 2001 approach that have been evaluated in comparison with the 2001 system. The aim is to develop the system to run more efficiently, have more control over the characteristics of those being imputed within missed households, and get 'close enough' to the age-sex calibration constraints to remove the need for the final 'Pruning and Grafting' stage.

3.1) Logistic rather than Multinomial Models

The system in 2001 used multinomial models to capture three (counted in both, counted by census only, counted by CCS only) of the four cells on a two-way table of census and CCS. The fourth cell (missed by both) was not included but the calibration of weights reflected this as the control totals were based on dual-system estimation. However, fitting a logistic model based on census response for CCS responders is effectively equivalent to modelling overall census coverage (what we would want to model) under the basic dual-system assumptions.

3.2) Modelling Missed Individuals

In 2001 we modelled individuals missed within counted households and as such we had no direct control on the individuals missed through missed households. For 2011 we propose modelling all the missed individuals in a single model but splitting the missed into its two components, in other words extending the logistic back to a multinomial but still based only on the CCS responders measuring census coverage. This allows us to easily calibrate the overall weights to recover the estimated totals at the individual level by the following key variables:

- age-sex at the LA level
- tenure, ethnicity, hard-to-count, and primary activity at the EA level

Once the overall weight has been calibrated it can be split into the two components and generate a weight for individuals missed within counted households and a weight for missed within missed households. A check at this stage will be necessary to ensure the weights for missed individuals in counted households do not create more households in any size category than our household size constraint allows but in the simulations this has not yet been implemented. The missed household modelling proceeds as in 2001.

However, we use a logistic model on the CCS responding households. At this stage we do not calibrate the household weights as they will be applied to the data after imputing the missed individuals within counted households.

3.3) Reversing the Imputation Order

In the 2001 system, there was no direct estimation of missed within counted households and missed from missed households. As such, it was considered preferable to carry-out the household imputation first, thereby having the most flexibility when imputing individuals as part of missed households, and use the within household imputation to fill-up as needed. This meant that household weights needed to recover household totals but did not always match well to the individual totals. As discussed earlier (section 2), this often resulted in imputing households that did not always contain quite the correct individuals and required the pruning and grafting stage to get it right.

The modelling proposed in section 3.2 gives direct control over the split between the two sources of under-count for individuals. This allows us to put the missed individuals in counted households into the database first, using the same algorithm as last time, knowing that we have controlled the number and characteristics of the individuals being imputed and accounting for those missed through missed households. Doing the within household imputation first essentially ensures that the counted households on the database are ‘complete’ with respect to the individuals we estimate they should contain. We can now apply the weights from the missed household modelling to the database. These weights can now be calibrated to recover household estimates by the following key variables:

- tenure, household size, and hard-to-count at the EA level,

The weights will also need to be calibrated to recover the key individual variables:

- age-sex at the LA level,
- primary activity last week and ethnicity at the EA level

The calibration approach ensures the weight for a household remains constant for all individuals within the household. This is achieved by treating each category of the individual variable as a household variable and counting how many individuals within the household fall into the category. Using these calibrated weights, we then run the household imputation system as last-time.

The household selected for imputing within the search category depends on the sort order within of the census file which is determined by specified variables and the household weights. Two sort orders are assessed in this paper. The first uses the 2001 variables (tenure, structure, hard-to-count stratum, ethnicity) with the addition of Local Authority and household size. This method is referred to as the 2011 base method. The second method of sorting adds age-sex groups to the base method variables, in order of coverage, so putting the 20-24 males and 20-24 females first, followed by young children and so on. This second method is referred to as the 2011 proposed method. The two sort orders are shown in Table 1.

To simplify the system, we do not search for a donor household but take the record that the system trips on, which is essentially a random event. This household is then placed based on the dummy form search as in 2001. In the few cases where random placement takes place, the imputed household is added to a postcode at random within the donor's ED.

Table 1: Sort order of variables used for 2011 base method and 2011 proposed method

2011 base method sort variables and order	**2011 proposed method sort variables and order**
Household tenure	Household tenure
Local Authority	Local Authority
	age-sex groups in order of coverage (M20to24, F20to24, F1to4, MF00, M1to4, M25to29, F25to29, M30to34, F10to14, M10to14, F30to34, MF5to9, MF35to39, MF15to19, MF45to49, MF55to59, MF50to54, MF40to44, MF70to74, MF60to64, MF65to69, MF80plus, MF75to79)
Household structure (collapsed)	Household structure
Hard-to-count stratum	Hard-to-count stratum
Household ethnicity	Household ethnicity
Household size	Household size

4) Results from initial Evaluations

As a baseline for comparison, a small number of simulations have been implemented using the final version of the 2001 system. The simulations are based on coverage models from the matched 2001 Census-CCS data that define coverage probabilities for households and individuals across the country for both the Census and the CCS. These are used to generate plausible pseudo-census and CCS data for groups of local authorities, referred to as Estimation Areas in 2001. Full details of the simulation structure can be found in Brown and Sexton (2009).

The system requires a set of calibration totals to work with such as the age-sex distribution. To allow us to concentrate just on the performance of the imputation system, we use the true totals as the calibration constraints, while the household and individual weights are based on applying the multinomial modelling to a simulated Census-CCS set of data. This is similar to the approach taken in the evaluations prior to 2001 (see Steele et al, 2002) and allows us to judge the additional variability introduced by the imputation system in return for bias reduction. The two performance measures are

$$\mathrm{RAB} = \frac{\frac{1}{10N}\sum_{e=1}^{N}\sum_{i=1}^{10}\left(T_{ei}^{(adj)} - T_e\right)}{\bar{\bar{T}}} \times 100 \tag{1}$$

$$\mathrm{RRAMSE} = \frac{\sqrt{\frac{1}{10N}\left\{\sum_{e=1}^{N}\sum_{i=1}^{10}\left(T_{ei}^{(adj)} - T_e\right)^2\right\}}}{\bar{\bar{T}}} \times 100 \tag{2}$$

where RAB (1) is the bias across 10 simulations at the ED level averaged across the EDs on a relative scale, while RRAMSE (2) is the mean square across 10 simulations at the ED level averaged across the EDs on a square-root relative scale. A small (zero) RAB represents correct placement averaged over the estimation area, while the RRAMSE shows how well we do at placing the imputed individuals / households at the ED level. (We get the right number of young men into the estimation area giving a zero RAB but the RRAMSE shows whether on average we are placing them in to the correct EDs.)

4.1) Performance of the System (computing)

The processing time of the system is important and any adjustments that we subsequently make should certainly not increase the time on the computer. In the simulations KO (an average coverage area in 2001) has run in around 2.5 hours with the last few minutes being used for the pruning and grafting. The revised approaches tend to run in a slightly shorter time overall but the actual imputation is quicker while the household calibration including individual constraints takes longer.

4.2) Performance of the System (statistical)

Looking at the basic structure of the database, in the first simulation 60.7% of the missed individuals were within missed households. The 2001 system added the correct number of individuals overall but 54.8% went in to missed households. The revised system including the full age-sex sort added close to the correct number of individuals (how close in the subsequent results) with 64% in missed households. This being closer to the 60.7% reflects the revised system trying to directly control the split.

Table 2: Overall performance at household level for the unadjusted census, the 2001 system and the 2011 systems

Method	RAB (%)	RRAMSE (%)	Shortfall
Unadjusted census	-3.93	5.21	7997
2001 system	0	0.90	0
2011 base system	0	0.77	-5
2011 proposed system	0	0.77	-1

We start by considering the performance at the household level. If we just consider the total number of households per ED, the Census has an RAB of -3.93% and an RRAMSE of 5.21%. The adjusted system reduces the RAB to zero by design (the system is calibrated to the true number of households by tenure and the pruning and grafting ensures this distribution is correct at the estimation area level) while the RRAMSE is only 0.90%. We get bias reduction across the estimation area and do well at placing missed households in the 'right' EDs. This is partly because in the simulation we create a 'perfect' set of dummy forms in the sense that each missed household is assigned a dummy form. The system does not use all the forms due to its matching rules but uses over 90% in each simulation so most households are placed in EDs with a missed household represented by a dummy form leading to the low RRAMSE.

The 2011 systems both produce a slightly reduced RRAMSE of 0.77%. The 2011 systems are putting in slightly too many households in some simulations. This is because the sorting methods sometimes result in many small weights being lower down the sorted census file which means that if too many households get imputed early on the weights never ‘catch up’. An option that we have considered is to sort the affected tenure classes so that the smaller weights are listed first in the census file. We have yet to establish the effect of this sorting on the quality of the imputation.

Turning to the performance by tenure, Figure 1 shows the RAB and RRAMSE. By design, the RAB is zero in the adjusted data but shows the expected negative bias in the unadjusted census for those renting. Across the categories, the adjusted data does better in terms of RRAMSE, particularly for those categories with the poor census coverage. For the part rent / part mortgage category both perform poorly but this group has only 1,207 households (less the one percent of households) in the estimation area and therefore missing single households, or placing single households in the wrong ED, will have (in relative terms) a large impact. There is little to choose between the performance of the three adjustment methods

Figure 1: Performance of the census and adjusted data in terms of RAB and RRAMSE for the household variable Tenure

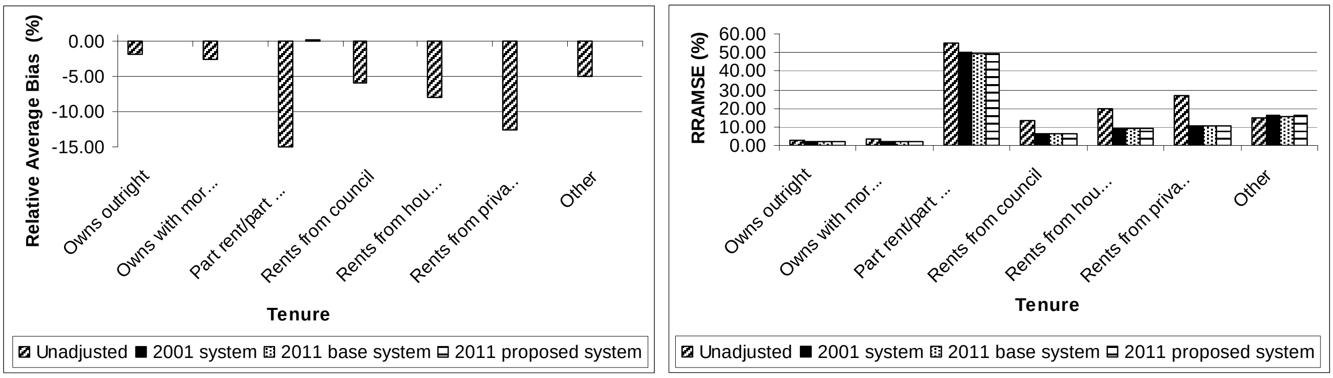


Table 3: Overall performance at person level for the unadjusted census, 2001 system and the 2011 systems.

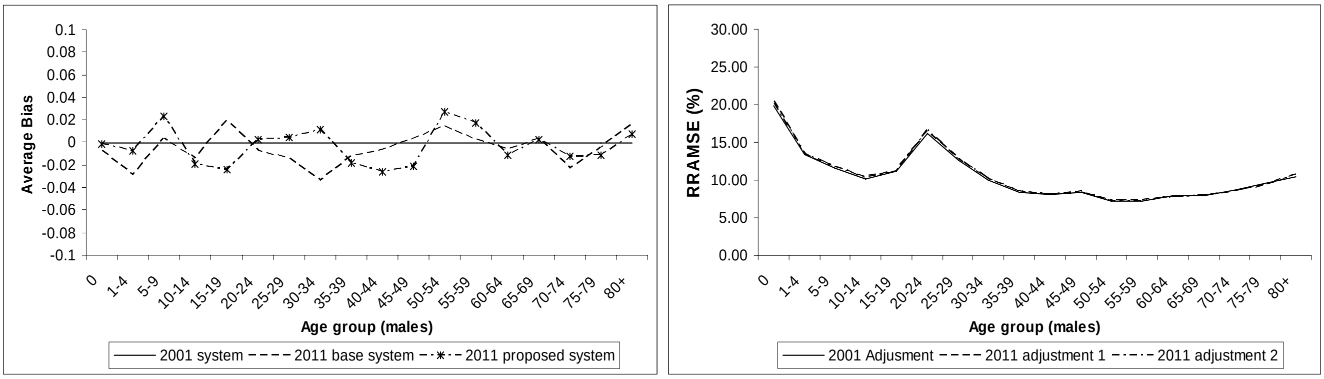
Method	RAB (%)	RRAMSE (%)	Shortfall
Unadjusted census	-6.08	7.74	29880
2001 system	0	2.89	0
2011 base system	-0.03	3.00	-139
2011 proposed system	-0.02	2.98	-78

We now turn our attention to the placement of individuals. If we just consider the total number of individuals per ED, the Census has an RAB of -6.08% and an RRAMSE of

7.74%. The 2001 system reduces the RAB to zero by design (the system is calibrated to the true number of individuals by age-sex and the pruning and grafting ensures this distribution is correct at the estimation area level) while the RRAMSE is reduced to 2.89%. We get bias reduction across the estimation area and do well at placing missed individuals in the 'right' EDs. At the ED level the average error in the adjusted database is less than three percent compared with over seven percent for the unadjusted census. The 2011 systems have a small overall relative average bias. This is because the 2011 systems do not calibrated exactly to the age-sex totals as there is no pruning and grafting done. In terms of RRAMSE the 2011 systems have qualitatively comparable performance to the 2001 system.

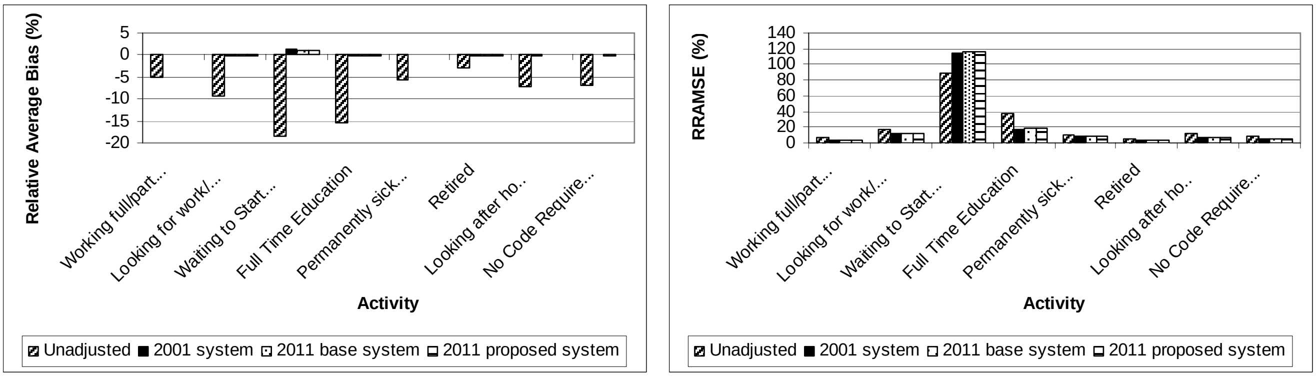
Figure 2 shows the average bias and RRAMSE for males by age-group for the 2001 system and the 2011 systems. (Similar but generally less extreme patterns are seen for females.) By design, the average bias at ED level is zero in the 2001 system. The 2011 systems vary across the age groups but stay well within plus or minus 0.05 persons on average. All three systems have similar performance for RRAMSE.

Figure 2: Performance of the adjusted data in terms of average bias (number of persons) and RRAMSE (%) for the individual variable Age-Group (males)



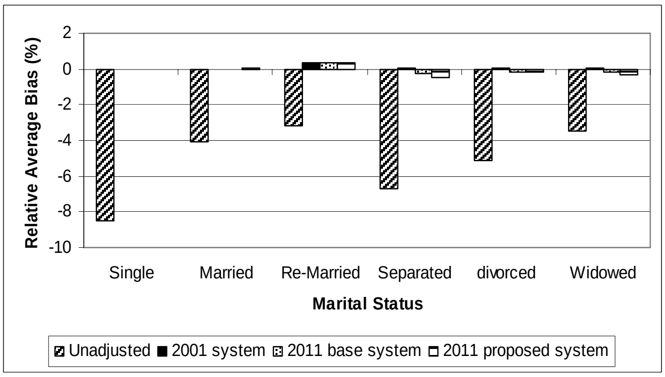
Age-Sex at the individual level is tightly controlled as the system calibrates to the variable and then is guaranteed to meet the calibration exactly in 2001 and get very close in the 2011 proposals. However, main economic activity is included in the model and calibration for individual within household coverage and the initial household weights are calibrated to the variable, but it is not guaranteed that the final database will precisely achieve the calibration constraint. Therefore, when we look at the RAB for the adjusted data in Figure 3, we see that it is very close for all three methods but not quite zero. However, we make very good improvement relative to the Census and in terms of the RRAMSE the adjusted data again is never worse than the Census (with the exception of ‘waiting to start work’) so the trade-off to reduce bias by introducing some variability in the placement at the ED level is working well. The ‘waiting to start work’ category contains only 297 individuals, less than 0.1% within the estimation area, so placing an individual within an ED will tend to have a big relative impact. The way the system works it will generally place people in EDs where it finds others with the characteristic. With a category this rare that means the system will add them to an ED where they are already counted but cannot place them in an ED where the only individual within the ED has been missed.

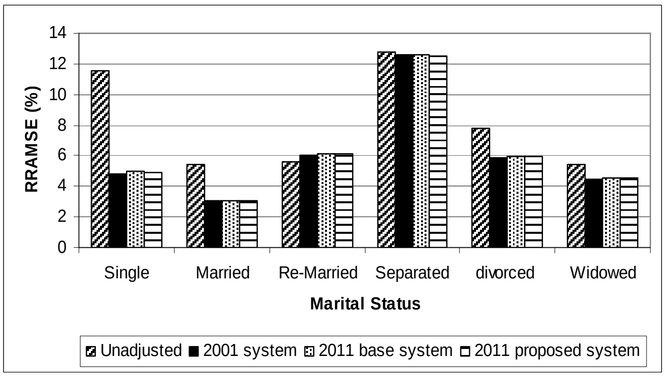
Figure 3: Performance of the Census and Adjusted data in terms of RAB and RRAMSE for the individual variable Primary Economic Activity



We also want to consider the performance of variables that the system has little direct control over. Marital Status at the individual level is partially accounted for through the size-type variable in the modelling but this variable is not calibrated so we would not necessarily expect the adjusted data to perform so well. When we look at the RAB for the adjusted data in Figure 4, we see that it is still very close but not quite zero. However, we continue to make very good improvement relative to the Census and in terms of the RRAMSE the adjusted data again is never worse than the Census. There is little to choose between the performance of the three adjustment methods in terms of RAB or RRAMSE,

Figure 4: Performance of the Census and Adjusted data in terms of RAB and RRAMSE for the individual variable Marital Status





5) The Way Forward

We now feel that the proposed system gets sufficiently close to the age-sex calibration without the need for pruning and grafting, and the system runs considerably quicker once the household weights are calibrated. These initial results show the performance of the proposed systems are certainly no worse than the 2001 system. Work will be undertaken to evaluate further estimation areas with more detailed analysis of variables within the database.

Further work is required to ensure the weights for missed individuals in counted households do not create more households in any size category than specified by the household size constraints.

Final development of the 2011 system still requires us to convert the code to:

- handle the structure of the 2011 variables,
- use OAs rather than EDs (this has been done but not integrated into the new system),
- include additional variables on migration, country of birth, and intention to stay.

These developments are happening as the code is developed to sit in the DSP system.

References

Abbott O. and Brown J. (2002) “Changes to the ONC Imputation System” ONC Steering Committee paper ONC(SC)02/01. Available at http://www.statistics.gov.uk/census2001/pdfs/sc0201.pdf

Abbott, O. and Brown, J. (2007). Coverage Adjustment Options for 2011: Early Ideas. Internal Paper. Available on request.

Brown, J. and Sexton, C. J. (2009) “Estimates from the Census and the Census Coverage Survey”, Paper presented at GSS Methodology Conference, London, June 2009. Available at http://www.ons.gov.uk/about/newsroom/events/14th-gss-methodology-conference--30-06-09/programme/index.html

Steele, F., Brown, J. and Chambers, R. (2002) “A controlled donor imputation system for a one-number census,” Journal of the Royal Statistical Society A, **165**. 495-522.