



ONS(ONC(SC))98/15

## **ONE NUMBER CENSUS STEERING COMMITTEE**

1. This paper presents a methodology for achieving the final stage of the One Number Census (ONC) process: the adjustment of the individual and household level databases to account for the estimated underenumeration.

2. Some results of a simulation study run by a prototype system are given, the remaining results will be presented at the meeting of the Steering Committee on 13<sup>th</sup> November.

**2. The Steering Committee are asked to:**

- a) **note the methodology**
- b) **approve development of the imputation approach, with weighting as a backup option**
- c) **provide any comments at the meeting on the 13 November 1998, or in writing by 27 November 1998.**

**Marie Cruddas  
Census Division  
Office for National Statistics  
Room 4200W  
Segensworth Road  
Titchfield  
Fareham  
HANTS  
PO15 5RR**

**October 1998**

## **IMPUTATION AND WEIGHTING EXECUTIVE SUMMARY**

The final stage of the ONC process is the adjustment of the individual and household level databases to account for the estimated underenumeration. At this stage there will be agreed population estimates, by age-sex distributions, for the national and Local Authority District (LAD) levels of geography. These estimates will be consistent with each other, so that LAD estimates will sum to national level estimates for each age-sex group.

This paper describes a methodology for cascading the population estimates down to the individual level. Within this there are two possible approaches: weighting and imputation. For practical reasons, largely relating to ease of handling outputs, the imputation approach is preferred and weighting is incorporated as an alternative within the same framework.

The imputation procedure is outlined below. The process is described in the paper in more detail and a description of the implementation of a prototype imputation system for simulated data is given. Results are given for imputation of household level characteristics for one of the simulated datasets, these show (see Table 4) that in general the method works well at recovering the 'true' distributions of the household characteristics, even for a variable not covered by the CCS. The remaining results, for individual imputation, will be presented at the meeting of the Steering Committee on 13<sup>th</sup> November 1998.

### **Step One – Modelling to Estimate Coverage Weights**

For areas in which the Census Coverage Survey (CCS) has been carried out the final stage of the ONC process starts by using matched Census and CCS data to model the probability of being counted in the Census in terms of the characteristics of individuals and households. This is possible in CCS areas where there are two 'independent' counts of the population. These models are applied to all individuals and households counted by the Census in order to calculate their "census coverage" probabilities. These probabilities in turn are inverted to form coverage weights.

For households the coverage weights are calibrated so that weighted counts of households found by the Census accord with the agreed estimates for the total population by household size for each LAD.

The modelling of census coverage underlying this procedure is based on the fact that there are two processes that cause individuals to be missed by the census. The first is when there is no contact with the household and therefore all the members are missed. The second is when contact with the household fails to enumerate all the members and therefore some individuals within counted households are missed. These two processes are treated separately by the methodology.

### **Step Two - Imputation of Completely Missed Households**

The second stage in the process uses the household weights to impute completely missed households. To do this households are grouped into "impute" classes defined by similar household characteristics and processed sequentially in order of increasing coverage weight. When the cumulated weighted count of the households gets more than 0.5 ahead of the

cumulated unweighted count a new household is imputed at or near the location where this event takes place. The donor household for this imputation is defined on the basis of the characteristics of the impute class as well as those households with the “current” weight and not only donates the household characteristics but all the individuals within the household as well. This process ensures that the total number of households after imputation matches the estimated LAD total. It will also match on totals defined by any other variables to which the household weights have been calibrated.

### **Step Three – Imputation of Individuals into Counted Households**

This is the most complex stage of the imputation as adding individuals to households naturally changes the structure of the recipient household. This stage is in two parts.

- The first part identifies how many individuals need to be imputed and obtains the appropriate donors. This is done in a similar way to the household imputation process described above. That is, individuals are processed sequentially in order of coverage weight within impute class. When the cumulated weighted count exceeds the cumulated unweighted count by more than 0.5 an individual needs to be imputed. The impute class and weight define the basic characteristics of that person. A donor household is then found that contains a person of the required type.
- The second part involves imputing the person into a “nearby” recipient household. The recipient household is the household nearest to the donor household in both space and household structure. When a suitable recipient is found the additional person from the donor household is imputed into the recipient household. Variables for individuals within the recipient household are edited so that the new household does not violate the census edit checks.

### **Step Four - Pruning and Grafting of Individuals**

The preceding stages of imputation add individuals to the census database, either as part of an imputed household or as additions to a counted household. Typically, this results in an “excess” of such synthetic individuals on the database. In addition, the number of households by variables such as tenure will be correct but the household size distribution for the LAD will not match the estimated distribution based on CCS data. The final stage of the imputation process therefore is to make sure that the totals of individuals match LAD totals by age and sex and that the household size distribution correct. A process of “pruning off” and “grafting on” of imputed individuals from the database is then carried out until these key LAD totals are achieved.

# **CREATING A ONC MICRO-LEVEL DATA BASE: IMPUTATION AND WEIGHTING OPTIONS**

**James Brown, Fiona Steel and Ray Chambers**

## **1 Introduction**

An integral part of the complete One Number Census (ONC) process is the adjustment of the individual and household level databases for the estimated underenumeration. It is this final stage that ensures consistency across all counts and tabulations produced from the census data. This paper describes two methodologies that have been proposed for this adjustment and the (ongoing) research into choosing between them. The first is weighting. This involves using the census data combined with the CCS data to produce coverage weights for different types of households and individuals. These weights are calibrated to agree with previously estimated population totals. The second is imputation. This takes weighting a stage further and uses the coverage weights in a donor imputation system to fill in the missing households and individuals on the database.

This paper concentrates on the second methodology as consultation suggests that this is the one favoured by the great majority of users. It is also the more complicated of the two methodologies as far as implementation is concerned. To a large extent the weights required for implementing a weighting approach are derived during the imputation process. The current proposal is to use an imputation based method to create the final ONC database subject to the results from the simulation study supporting this. Weighting is the second option that can be used if imputation does not perform satisfactorily.

## **2 Imputation Methodology**

The starting point for the methodology is the assumption that, following the CCS, estimates have been made of the population by age and sex down to LAD level. This relates to stages one to five laid out in the ONC Methodology Paper ONS(ONC(SC))98/16. Two caveats are necessary. First, this research assumes perfect matching. The sensitivity of this research to the levels of imperfect matching suggested by the matching research will be part of the next stage of the imputation research. Second, for each LAD it is also assumed that estimates of the true number of households are available by tenure, and possibly other characteristics. This estimate of the total number of households is consistent to the age-sex totals as defined by the distribution of households by size.

From this starting point there are a series of steps to an adjusted database.

- i) Use the data to model undercount and calculate coverage weights, first for households and then for individuals in counted households.
- ii) Use the household weights in a donor imputation system to impute households completely missed by the census.
- iii) Use the individual weights to impute individuals missed by the census in counted households.

- iv) Check that the necessary constraints at the LAD level have been met. This is achieved by pruning out imputed individuals or grafting on individuals from the database and ensures that the database satisfies the consistency requirements for a ONC.

The methodology for each step is laid out in the following sections.

## 2.1 Step One – Modelling to Estimate Coverage Weights

This first step is required whether an imputation or a weighting methodology is used. The method proposed distinguishes between two ways in which individuals are missed. The first is when the whole household is missed. By definition all individuals within that household are missed with probability one. The second is when contact is made with a household but not all the members are counted.

### 2.1.1 Derivation of Household Coverage Weights

Following the census and the CCS all households within CCS areas can be fitted into one of the following categories:

- 1) Counted in the Census, but missed by the CCS
- 2) Counted in the CCS, but missed by the Census
- 3) Counted in both the Census and the CCS
- 4) Missed in both the Census and the CCS

A simplifying assumption is that category four contains no households, that is no household is missed by both. While an unrealistic assumption, the households missed by both are accounted for in the dual system estimates at the design level and the final imputed database is constrained to satisfy the totals at the CCS design area. Excluding category 4, the categories above define a multinomial outcome variable with categories 1,2,3 from above that can be modelled for each LAD using:

$$\log\left(\frac{\theta_{jkel}^{(t)}}{\theta_{jkel}^{(3)}}\right) = \lambda_1^{(t)} Z_{jkel} \quad t = 1,2 \quad (1)$$

where  $\theta_{jkel}^{(t)}$  is the probability that household  $j$  in postcode  $k$  in ED  $e$  in LAD  $l$  with household level characteristics defined by  $Z_{jkel}$  is in category  $t$ . (Model (1) uses category 3 as the reference category.) With matched data from the census and CCS, this model is straightforward to fit. However, it only accounts for variability between areas based on the observed characteristics of the households. Therefore, the model can be extended to include random effect terms to estimate the additional postcode and ED variability not captured by the variables in  $Z$ .

The model fitted in the CCS areas is used to estimate underenumeration in the non-CCS areas. This gives predicted coverage probabilities for all households. The prediction is straightforward in the case of fixed effect modelling. However, when random effects modelling is used it is not straightforward to identify which values should be used for higher level residuals in non-CCS areas. The independence assumption traditionally made in multilevel modelling leads to estimates of zero for postcode and ED random effects which implies that there is no unobserved variability at higher levels across the non-sampled areas.

This may not be altogether realistic and another possibility is to smooth the observed residuals spatially. This proposal was initially investigated in Brown *et al* (1998) but further work is needed to assess any gains in prediction from the model against the additional complexity of fitting the models. In this paper a fixed effects model is used.

The probabilities for each response category estimated under model (1) define a coverage weight for each household (h/h) counted in the census

$$W_{jkel}^{h/h} = \frac{1}{\theta_{jkel}^{(1)} + \theta_{jkel}^{(3)}}$$

that can be applied to each household in the household database. However, the resulting weighted sum of counted households will not, in general, match the totals estimated for the LAD. This will require the weights to be calibrated to the LAD marginal totals for key household variables, such as tenure, using iterative scaling. This involves scaling the weights to match the marginal totals for one variable, then scaling to the next variable, and so on until the weights converge and all the weighted totals agree with the relevant marginal totals. An example is given in Appendix I.

### 2.1.2 Derivation of Individual Coverage Weights

To calculate coverage weights for those individuals counted within counted households, two assumptions are necessary regarding coverage of individuals in CCS areas.

- a) If a household is only counted by the census and not by the CCS, then no individuals from that household are missed by the census.
- b) Similarly if the CCS counts the household and the census does not, then no individual from that household is missed by the CCS.

These assumptions are necessary because a household counted by only one source has no second list against which counted individuals can be compared. Although this assumption does not hold in general, people missed as a consequence are accounted for through constraining to population totals at the LAD level. In this case the possible categories of counted individuals are:

- 11) Counted in the Census, but missed by the CCS
- 12) Counted in the CCS, but missed by the Census
- 13) Counted in both the Census and the CCS

These categories are then used to define the response in the following multinomial model:

$$\log\left(\frac{\pi_{ijkl}^{(r)}}{\pi_{ijkl}^{(13)}}\right) = \beta_1^{(r)} X_{ijkl} + \gamma_1^{(r)} Z_{ijkl} \quad r = 11, 12 \quad (2)$$

where  $\pi_{ijkl}^{(r)}$  is the probability that individual  $i$  in household  $j$  in postcode  $k$  in ED  $e$  in LAD  $l$  with individual characteristics defined by  $X_{ijkl}$  and household characteristics defined by  $Z_{ijkl}$

is in category  $r$ . (Model (2) uses category 13 as the reference category.) As before this model can also be extended to include random effects terms.

As with the household model the fitted model is then used to extrapolate to non-CCS areas and to give predicted coverage probabilities for all individuals missed within counted households. Again, this is straightforward for the fixed effects model but more work is needed if random effects and spatial smoothing are to be used. The probabilities estimated under the model then define a coverage weight for each individual ( $ind$ )

$$W_{ijkl}^{ind} = \frac{1}{\pi_{ijkl}^{(11)} + \pi_{ijkl}^{(13)}}$$

that can be applied to the individual database. As before the resulting weighted sum of census counted individuals will not be equal to the LAD totals but this is allowed for in step four.

## 2.2 Step Two – Imputation of Completely Missed Households

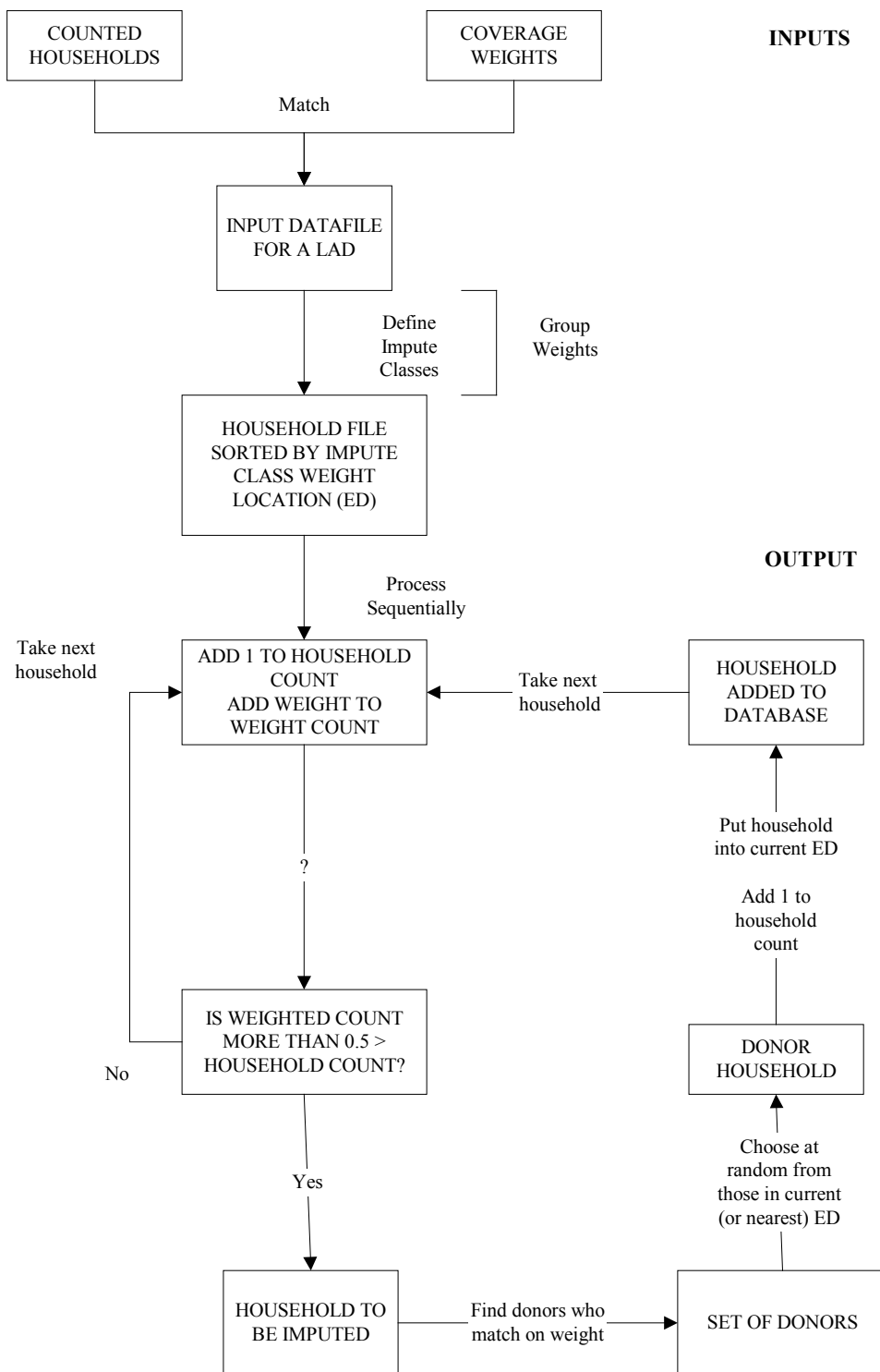
The methodology for the imputation of individuals in households that were missed by the census is outlined in Figure 1. Initially the household-based file of counted households in a LAD is matched to the file of calibrated household coverage weights (as described in 2.1.1). Households are grouped into impute classes in such a way that households within a class are homogeneous with respect to household characteristics. These will be largely defined by the characteristics on which the household coverage weights are based. Within the impute classes the file is sorted by coverage weight, then by ED and postcode. For simplicity, in the simulation study the impute classes are constructed by grouping coverage weights into bands containing approximately 10,000 households. The processing block is an impute class within an LAD.

Within each processing block, households are processed sequentially and running totals are kept of the unweighted household count and the weighted household count (calculated using calibrated coverage weights). Whenever the weighted count exceeds the unweighted count by more than 0.5, households are imputed into the ED currently being processed until the difference between those running counts is less than or equal to 0.5. When a household is imputed, the unweighted household count is incremented by 1 and the imputed household is assigned a zero household coverage weight. This step continues until all counted households in the processing block have been processed.

In order to assign characteristics to the imputed households, a donor imputation method is used. For a given imputed household, a donor is selected at random from among the counted households with the same weight and in the same ED as the counted household that was processed immediately before the imputation. Once a donor has been selected, the characteristics of the household and its occupants are copied to the imputed household. The imputed household is then assigned at random to a postcode within the ED.

At this stage dummy forms (a dummy form is completed by the enumerator for an identified non-responding household) have not been utilised and in effect the missed households are being spread throughout the EDs. However, this is not a rejection of dummy forms but a recognition that simulating them would be very difficult. Clearly in 2001 it makes sense to utilise the dummy forms when choosing the location of imputed households and to locate households in EDs where dummy forms have been completed by enumerators.

**Figure 1. Step Two – Household Imputation**



## 2.3 Step Three – Imputation of Individuals into Counted Households

Figure 2 outlines the methodology used for the imputation of individuals into counted households. The individual-based file of counted individuals is first matched to the file of uncalibrated individual coverage weights (as described in 2.1.2). As in the household imputation impute classes are constructed, defined by the characteristics on which the individual coverage weights are based. These are then sorted by weight and geographical location. In the simulation study impute classes are formed by grouping coverage weights into bands. Within a processing block (impute class within a LAD), counted individuals are processed sequentially. When the weighted count of individuals exceeds the unweighted count by more than 0.5, individuals are imputed in the current ED until the difference is less than or equal to 0.5.

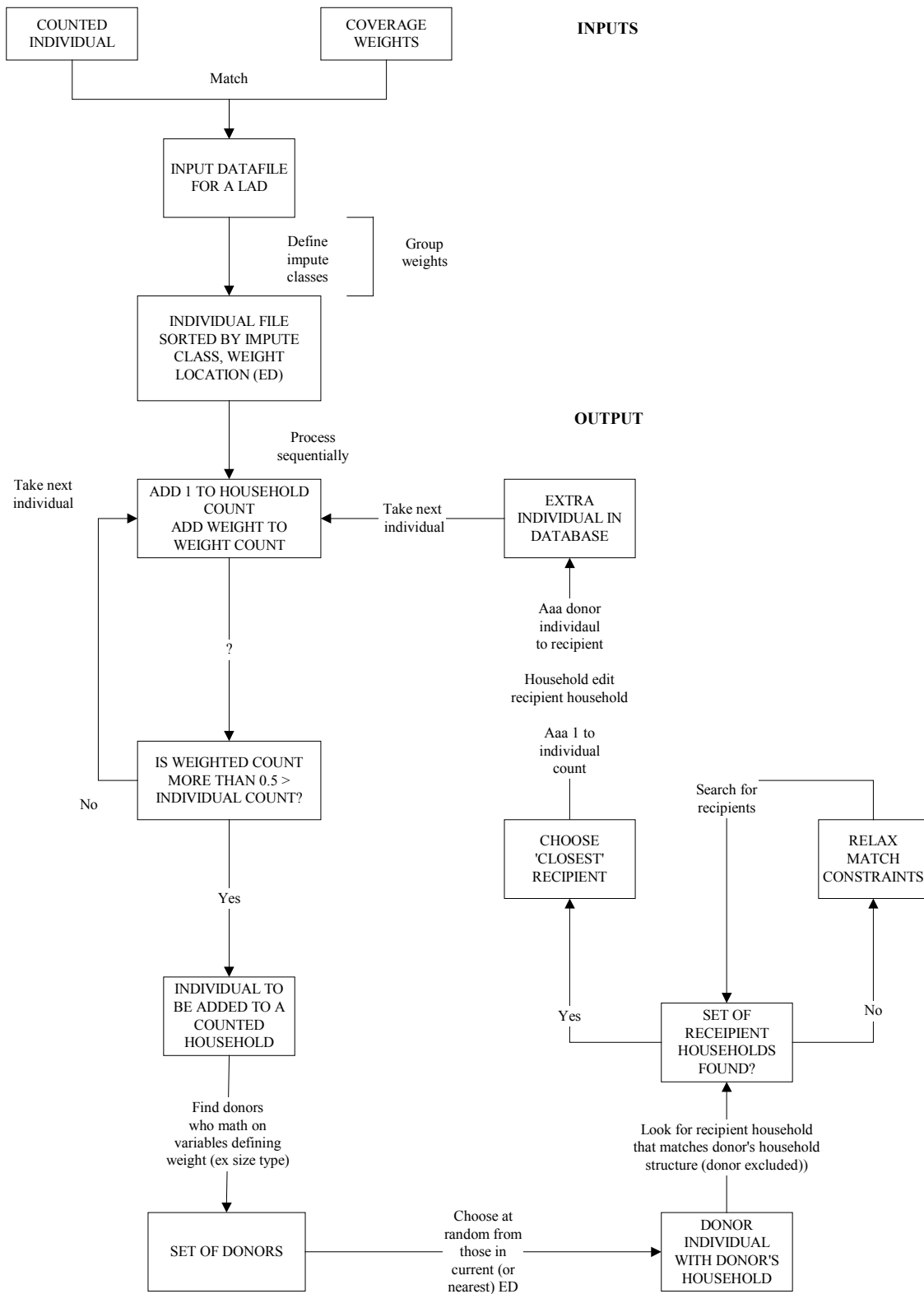
Several steps are necessary to assign characteristics to the imputed individuals. Some of an imputed individual's characteristics are determined by the weight of the last counted individual that was processed before the imputation. The remaining individual characteristics are copied from a suitable donor. Finally, the LAD is searched for a suitable recipient household in which to place the imputed individual. The household characteristics for an imputed individual come from the selected recipient. The donor and recipient searches are now described in more detail.

### 2.3.1 The Search for Donor Individuals

The type of donor sought depends on the characteristics of the individual to be imputed (as determined by the coverage weights). In the simulation study that follows choice of donor depends on the age, marital status and household structure of the individual being processed when the weighted count suggests that an individual needs to be imputed. The household structure variable is derived from the census, i.e. it refers to the household *without* the individual that needs to be imputed. In the simulation study, household structure is categorised as follows: 1) single person, 2) single parent with all children under 16, 3) married couple, 4) married couple with all children under 16, 5) unrelated adults, and 6) mixed (including families with children aged 16 or over).

To illustrate the donor search, consider two individuals that the model suggests need to be imputed. The first individual is a single (never married) person aged less than 16 who is missing from a household whose structure was defined as 'couple' at the census. The second is a single (divorced/widowed) person who is missing from a household defined as 'unrelated adults' at the census. The donor for the first individual must also be single (never married) and aged less than 16, but the structure of their household will be 'couple with children under 16' and the household containing the donor will be of size three. For the second individual, a suitable donor would be a divorced or widowed person living in a household with unrelated adults. In addition, potential donors should match on a previously specified set of other individual and household characteristics. In the simulation that follows these are age (categorised into 12 groups), sex, economic activity, tenure and hard to count (HtC) index. Finally, some restrictions regarding the size of the donor household are introduced. These depend on the imputed individual's household structure and are necessary to enable recipients to be found in the next stage (see 2.3.2). The household that contains the first donor illustrates this in the above examples. In that case the household must be of size three.

**Figure 2. Step Three – Individual Imputation**



Where more than one suitable donor is found, a donor is selected at random from those in the ED closest to the ED in which the imputed individual was initially placed. If no donor is found, the matching criteria are relaxed so that an exact match on variables such as HtC index, tenure and household ethnicity is no longer required. The number of matches on these variables is calculated for each potential donor and the donor is selected at random from the closest matches.

Once a suitable donor is selected, their individual characteristics become the characteristics of the individual to be imputed.

### **2.3.2 The Search for Recipient Households**

The next stage of the individual imputation involves searching for a suitable household into which the imputed individual can be placed. The other occupants of the donor's household are first considered and then a search is done for a recipient household of an appropriate household structure (with one less individual than the donor's household) whose occupants match those of the donor's household on age and sex. The potential recipients are then ordered according to the number of matches on other household characteristics. At the moment these are tenure, HtC index and household ethnicity. If more than one of the potential recipients match on the same number of household characteristics, one is selected at random from those in the ED nearest to the ED being processed when it became necessary to impute an individual.

To illustrate both searches, consider the following example. Suppose the individual to be imputed is a male aged 0-4 from a household defined as 'couple' at the census. In other words the weights are suggesting that couples with just one child, in this case a baby boy, have missed that child from the census form. Therefore, households that contain suitable donors must be of structure 'couple with children under 16' and of size three. Suppose the occupants of the chosen household containing the donor were of the following age and sex:

1. Female aged 25-29
2. Male aged 30-34
3. Male aged 0-4 (the donor)

To find a suitable recipient household for the imputed child it is necessary to search for a two-person household of structure 'couple' with a female aged 25-29 and a male aged 30-34 that matches the donor household and individuals on as many household and individual characteristics as possible.

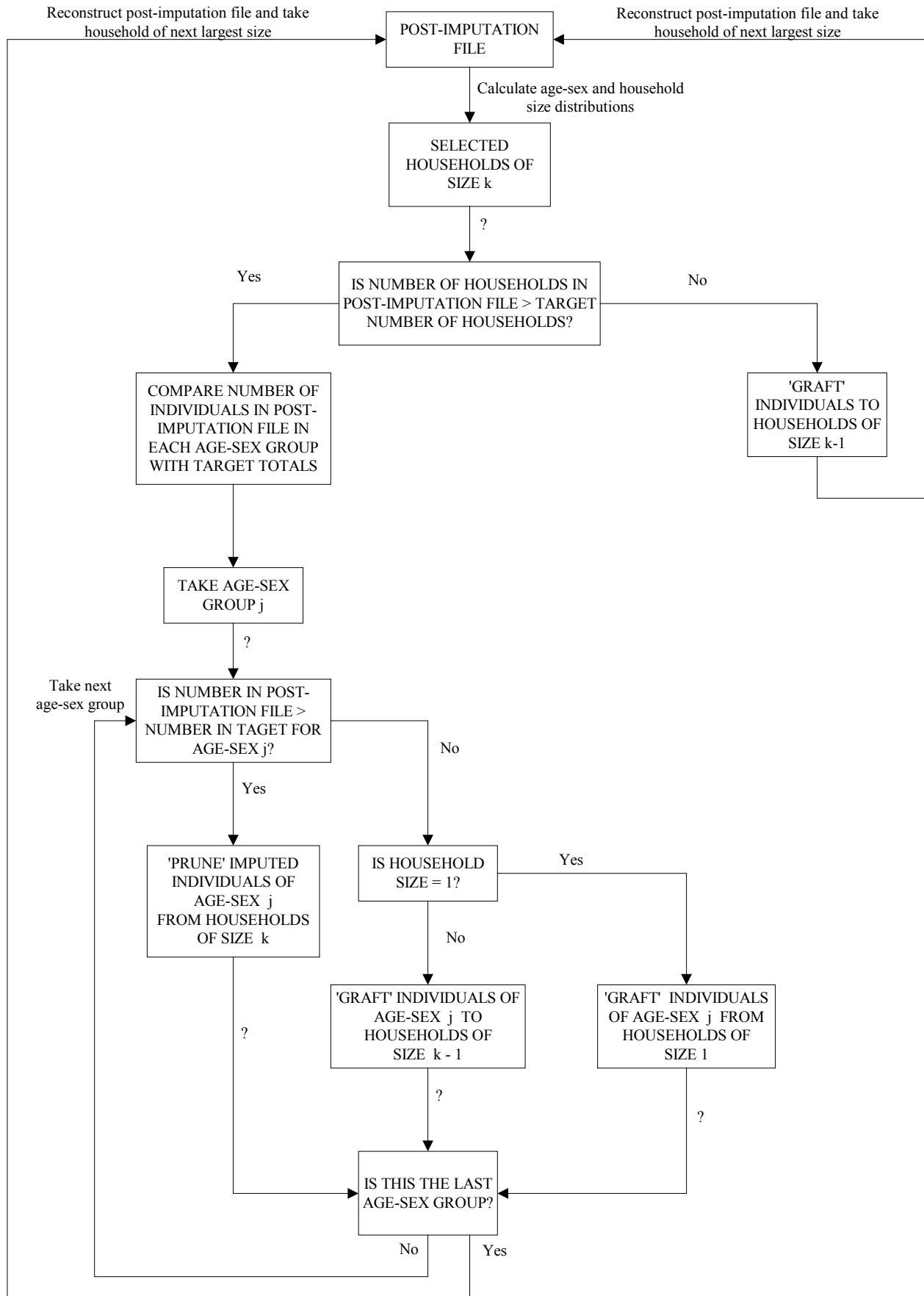
If no recipients are found, the matching criteria are relaxed as follows. First, the size of the recipient household is no longer required to equal the donor's household size minus one, unless the household structure constraints demand this. Instead, the match is made on the number of adults but not on the number of children. If there are children in the donor's household (not counting the donor), however, the recipient household must also have children. Further, recipient households where the ages of the adult occupants are within one age group of adults in the recipient household are accepted, provided they are of the correct sex. Finally, it is not required that children in the recipient household match children in the donor's household on either age or sex. Recipient households are still required to be of the same household structure as the imputed individual.

The imputed individual is now placed into the selected recipient household and takes their household characteristics. In some cases this will alter the household structure of the recipient household.

#### **2.4 Step Four – Pruning and Grafting**

Although Steps 2 and 3 above will lead to the correct number of households, due to the calibration of the household weights, with the correct distribution for the household variables to which household weights are calibrated, the household size distribution will be incorrect. This is due to individuals being imputed in both Step 2 and Step 3 that leads, in general, to too many larger households. To ensure that the household size distributions and age-sex distributions derived from the ONC database agree with the ONC estimates of their distributions at the LAD level, some addition and/or deletion of imputed individuals from imputed and counted households will be necessary. The basic idea of this procedure is to start at the largest households and work down to households of size one, adding and deleting people to move households up or down in size. This process is best described by use of an illustrative example. Figure 3 gives a flow chart of the process.

**Figure 3. Step Four – Pruning and Grafting**



Consider the following example, assuming for simplicity that there are four age-sex categories with a maximum household size of six. Suppose the age-sex by household size distribution in the individual file after both imputation steps, called the post imputation file, is as follows (the target totals from the ‘true’ distributions are shaded):

*Table 1: Illustration of household size by age-sex distribution for a hypothetical post imputation file*

		Household size						(# indiv)	<b>(True # indiv)</b>
		1	2	3	4	5	6		
Age-sex group	1	2	3	4	4	6	11	(30)	<b>(40)</b>
	2	2	2	8	10	15	17	(54)	<b>(45)</b>
	3	2	7	8	14	14	18	(63)	<b>(38)</b>
	4	4	4	10	16	20	14	(68)	<b>(49)</b>
# households		10	8	10	11	11	10		
(# individuals)		(10)	(16)	(30)	(44)	(55)	(60)	(215)	<b>(172)</b>
<b>True # h/hs</b>		<b>15</b>	<b>15</b>	<b>10</b>	<b>8</b>	<b>7</b>	<b>5</b>		
<b>(True # indiv)</b>		<b>(15)</b>	<b>(30)</b>	<b>(30)</b>	<b>(32)</b>	<b>(35)</b>	<b>(30)</b>		

In this case, there are  $215-172=43$  extra individuals who need to be pruned. Comparing the observed number of individuals with the true number in each age-sex group; there are

- 10 too few* in age-sex group 1
- 9 too many* in age-sex group 2
- 25 too many* in age-sex group 3
- 19 too many* in age-sex group 4

Starting at the largest household size (six here), the observed number of households is compared with the true number of households. There are five extra households of size six that need to be pruned. Five individuals from households of size six are deleted in proportion to the overall requirements as follows:

- Add*  $5 \times (10/43) = 1.16 \approx 1$  from age-sex group 1
- Delete*  $5 \times (9/43) = 1.05 \approx 1$  from age-sex group 2
- Delete*  $5 \times (25/43) = 2.91 \approx 3$  from age-sex group 3
- Delete*  $5 \times (19/43) = 2.21 \approx 2$  from age-sex group 4

To add an individual from age-sex group 1 (grafting), a household of size six is selected that contains an individual from age-sex group 1. From the selected household, one individual from age-sex group 1 is chosen as the donor. The search for a recipient household is then carried out amongst five person households whose occupants match the occupants of the donor’s household (not counting the donor) on age and sex and as many household

characteristics as possible. One such household is selected at random. In some cases, especially for large household sizes, it may not be possible to find a suitable donor and recipient household. In such situations, the matching criteria are relaxed as in the donor and recipient searches described in sections 2.3.1 and 2.3.2. This five person household is then made into a six person household by grafting the donor into it. Next, one imputed individual of age-sex group 2, three of age-sex group 3 and two of age-sex group 4, are deleted from households of size six.

The revised age-sex by household size distribution is now:

*Table 2: Household size by age-sex distribution for a post imputation file after stage 1 of pruning and grafting*

		Household size						(# indiv)	<b>(True # indiv)</b>
		1	2	3	4	5	6		
Age-sex group	1	2	3	4	4	?	?	(31)	<b>(40)</b>
	2	2	2	8	10	?	?	(53)	<b>(45)</b>
	3	2	7	8	14	?	?	(60)	<b>(38)</b>
	4	4	4	10	16	?	?	(66)	<b>(49)</b>
# households		10	8	10	11	16	5		
(# individuals)		(10)	(16)	(30)	(44)	(80)	(30)	(210)	<b>(172)</b>
<b>True # h/hs</b>		<b>15</b>	<b>15</b>	<b>10</b>	<b>8</b>	<b>7</b>	<b>5</b>		
<b>(True # indiv)</b>		<b>(15)</b>	<b>(30)</b>	<b>(30)</b>	<b>(32)</b>	<b>(35)</b>	<b>(30)</b>		

There are now  $210-172=38$  extra individuals (9 too few in age-sex group 1, 8 too many in age-sex group 2, 22 too many in age-sex group 3 and 17 too many in age-sex group 4).

The number of households of size six is now correct and the algorithm moves on to households of size five. There are nine households too many of this size. Therefore nine individuals need to be delete from five person households through a mixture of pruning and grafting. The number of individuals that are pruned or grafted from each age-sex group depends on the overall number of individuals that need to be added/deleted from each.

The process is then repeated for households of size four, three, two and one, or until the household size and age-sex distributions are consistent with the target totals.

In general, it is possible that for some household sizes, there will be fewer households in the marginal totals generated from the post imputation file than there are in the estimated true marginal totals. In this case, it is necessary to graft on individuals to smaller households and therefore create larger households. For example, suppose the post imputation file started with five too few households of size six. Five individuals from the post imputation file would be chosen and grafted into five households of size five. This would give the required extra five

households of size six, but five fewer households of size five. If this led to there being too few households of size five, it would be necessary to repeat this procedure to construct more households of size five.

## **2.5 Maintaining Marital Status Structures within Households**

The final step of the whole process to generate the ONC database is to ensure that married imputed individuals have a married partner in their recipient household. The simulation study assumes that at the editing stage the marital status of individuals who are reported as married is changed to single if no married partner is counted in the census. Thus when a married individual is imputed into a household there will be no married partner. A further step is required to find a suitable partner for married imputed individuals (the person nearest in age and of the opposite sex) and to change their marital status accordingly.

## **3 Weighting Methodology**

The application of weighting, as opposed to imputation, to produce a ONC database requires much less computation but has other disadvantages. This section briefly describes the methodology proposed for weighting and mentions some of the issues. If weighting is chosen as the preferred option for a ONC database then more work will be needed.

There are three basic stages. After the modelling of both households and individuals an overall coverage weight for individuals is calculated. This accounts for individuals being missed at both stages. A household coverage weight is also produced as before. The second stage involves calibrating the weights. First, the individual and household weights need to be calibrated to control totals for the marginal distributions of individual and household variables at the LAD level. Second, the individual and household weights need to be calibrated to each other. This ensures consistent answers from the two sets of weights. However, this calibration cannot be done for every possible variable and therefore there may be some household and individual tabulations that would not be consistent. For example, the implied number of individuals by tenure from a household tabulation of tenure by household size may not be consistent with the distribution of tenure on an individual tabulation unless the two sets of weights have been calibrated on tenure. The final stage uses the weighted versus the unweighted count idea from the imputation methodology to ‘randomly’ round some of the weights up to two, at the points where an individual or household would be imputed, while the rest are rounded to one. This ensures integer counts on tabulations that are consistent with the calibration constraints.

## **4 Simulation Study**

The aim of the simulation study is to evaluate both methodologies. While imputation is popular with users it is useful to highlight the possible risks of such a black box approach and compare this with the results from weighting.

### **4.1 Generation of the Data**

As with any simulation study the exact nature of the results will depend on the way the data have been simulated. For this study ten censuses have been generated from an LAD of 1991

Census records. In this case the true LAD population size is 445,267, in 171,206 households. Further details of the LAD used and the CCS design applied to it are given in Appendix II.

Censuses have been generated in a slightly different manner to previous simulations in the ONC Project. Each household has a probability of being counted. This varies according to the HtC index, the tenure of the household, and the size of the household, as well as the characteristics of the individuals within the household. Private rented and small households are more likely to be missed. The first stage of the census scans the household file and takes independent binomial trials on each household to decide if it is counted. Each individual also has a probability of being counted that varies according to age, sex, the HtC index, and economic status. The second stage of the census scans the individual file for those in the households counted at stage one. Independent binomial trials are then taken to establish whether the individual within the household is counted. A final stage removes households where all adults have been missed. For the CCS the probability of an individual or household being counted depends mainly on the coverage set for the CCS. However, non-response increases slightly for small households and for individuals it varies by age and sex.

As the aim is to evaluate the imputation procedure, the census and CCS have been generated independently of each other. The implications of breaking this assumption will be tested at a later stage, as this will impact on the accuracy of any estimated control totals. The overall census coverage has been set at 95 per cent, but this drops below 90 per cent for the young men. For each census a CCS has been simulated with a 90 per cent coverage of households and 98 per cent coverage of individuals within counted households. This second value is high but it is expected that with well-trained interviewers people should not get missed once contact has been made.

For modelling purposes some additional variables have been calculated based on each simulated census. Household structure, as described in section 2.3.1, is calculated for each counted household based on those individuals observed in the census. It is important to remember that the missed individuals will cause the observed structure of a household to change from census to census. There is also a household ethnicity variable, a simplified tenure variable, an economic status variable, and a marital status variable. These are all defined in Appendix III.

## **4.2 Generation of the Coverage Weights**

For each of the ten simulated data sets three multinomial models have been estimated, one for the missed households based on the model in section 2.1.1, one for missed adults and one for missed children that are both based on the model in section 2.2.2. The explanatory variables used for the household models are tenure, household ethnicity, household structure, and the enumeration district's HtC index. For missed individuals within counted households children have been considered in a separate model, as they do not have an economic status (as measured by the census). The explanatory variables used in the model for children are sex, age group at the individual level, a simplified tenure variable and the number of counted adults based on the household structure variable at the household level, along with the enumeration district's HtC index. The model for adults extends the model for children to include economic status, marital status at the individual level and the full household structure variable at the household level. It is important to remember that census data are always used when these are available. CCS data are only used for households completely missed by the

census and the individual characteristics of individuals missed by the census in counted households but not their household characteristics.

All the fitted models have estimated main effects only. The use of interaction terms will be investigated as the research continues. In addition only fixed effects models have been estimated at this stage. It is intended that at a later stage the results can be compared to models with random effects. For each fitted model a set of coverage weights have been calculated. There are some large weights generated, particularly for the missed individuals. In general this is not a problem as the large weights are associated with combinations of characteristics that are extremely unlikely in practice, for example retired twenty year olds. These people do not usually exist and are effectively structural zeroes in the data. It is proposed that judgement is used to eliminate the chance of over imputing such people.

The final stage is the calibration of the household weights to satisfy marginal distributions estimated at the local authority district level. For this simulation the ‘true’ marginal distributions have been used, as the aim here is to test the imputation rather than the ability to estimate totals at a higher level. Clearly this estimation will effect the overall accuracy of a ONC database and this will be investigated as the research develops. The weights have been calibrated to their true distribution by tenure, household ethnicity, and HtC index. Using the HtC index ensures that, in general, the hardest to count enumeration districts will get more imputed households. The calibration was done using an iterative scaling algorithm that converged very rapidly.

### 4.3 Household Imputation

For Step 2, households in the census file are grouped into impute classes as follows. The household file is first sorted by calibrated household weight, then households are processed sequentially. If the household currently being processed has a different weight from the last processed household, and the number of households in the current impute class is greater than 10,000, then a new impute class is formed. This generates 15 impute classes. After combining imputed households across classes, 6,017 were imputed to give a total of 171,205 households. This stage of the imputation procedure is very fast and for the simulated data takes less than one hour to process. Table 3 shows the distribution of households in the true, imputed, and unadjusted census databases.

*Table 3: Distribution of Households by Tenure*

Tenure Categories	True Distribution	Imputed Distribution	Census Distribution
Owner occupied – mortgage	78,421	78,425	76,542
Owner occupied – outright	43,968	43,964	43,108
With job, farm, shop	1,805	1,801	1,749
Local authority	27,848	27,851	26,709
New town corporation	416	418	392
Housing association	6,039	6,038	5,715
Private landlord – furnished	5,689	5,688	4,667
Private landlord – unfurnished	7,020	7,020	6,306
TOTAL	171,206	171,205	165,188

Table 3 clearly demonstrates the impact of calibrating the household weights to the estimated (or in this case the true) tenure distribution for the local authority district. The calibration ensures that not only is the overall undercount of households corrected but it's differential nature is also corrected. However, the marginal distribution for the imputed database is not an exact match. It is likely that this is a result of some household weights being less than one (effectively suggesting an over count of a particular type of household) and therefore the weighted count is sometimes behind the unweighted count.

To evaluate the performance of the household imputation further, chi-square tests are used to compare the census and post imputation distributions of uncalibrated household characteristics with their true distributions. The results from this initial evaluation are shown in Table 4.

*Table 4: Initial Results - Evaluation of Household Imputation*

Variable	Truth	Imputed	Census
Building type			
Caravan	282	276	268
Detached	17906	17989	17640
Semi-detached	61878	62104	60657
Terraced (including end)	65195	65081	62542
Purpose built flat (commercial block)	1956	1956	1834
Purpose built flat (in block of flats)	19771	19704	18705
Converted house/flat (own entrance)	1125	1136	1050
Converted house/flat (shared entrance)	3093	2959	2492
Chi-squared test (verses true distribution)*		7.68 (0.362)	116.11 (0.000)
Multi-occupied identifier			
Not multi-occupied	166988	167110	161646
First household (multi-occupied)	1281	1259	1107
Subsequent households (multi-occupied)	1798	1719	1452
Last household (multi-occupied)	1139	1117	983
Chi-squared test (verses true distribution)*		4.36 (0.225)	73.52 (0.000)
Type of accommodation			
Not applicable	166988	167110	161646
One roomed flat	268	245	208
One room or bedsit (not self-contained)	1178	1104	875
Self-contained with 2+ rooms	2446	2449	2204
Not self-contained with 2+ rooms	326	297	255
Chi-squared test (verses true distribution)*		9.29 (0.054)	93.41 (0.000)
Number of cars			
0 cars	69662	69384	66037
1 car	70706	70837	68805
2 cars	26064	26178	25643
3+ cars	4774	4806	4703
Chi-squared test (verses true distribution)*		2.07 (0.559)	37.38 (0.000)

\* For the Chi-squared tests the figures in brackets show the significance level of the test of equality of distribution between the true and imputed or census estimates.

The chi-squared results in Table 4 clearly show that not only is the census distribution wrong in terms of numbers but the shape of the distribution is also significantly different from the truth. However, after imputing households not only are the numbers much nearer the true distribution in absolute terms but the shape of the distribution is now also much closer to the truth. For three out of the four variables the p-values for the chi-squared test are much greater than any standard significance level. For the variable ‘type of accommodation’ the distribution on the imputed database is marginally significantly different from the truth although this difference is much less significant than for the census database. These initial results are impressive and demonstrate that, at the local authority district, with good control totals, the household imputation is a large improvement on the census.

#### **4.4 Individual Imputation**

In the step to impute individuals into counted households, individuals were first grouped into impute classes. The procedure used to define impute classes for individual imputation is similar to that used for household imputation. Individuals are sorted by their coverage weight before being processed sequentially. If the weight of an individual is different from that of the last processed individual and the number of individuals in the current impute class exceeds 20,000, then a new class is formed. This leads to 21 impute classes. A total of 12,719 individuals were imputed.

##### **4.4.1 The Search for Donor Individuals**

With exact matching on age group, sex, economic activity, tenure and HtC index, it was possible to find donors of an appropriate household structure and household size for all but around 200 imputed individuals. Donors were found for the remaining individuals after relaxing the matching criteria as described in 2.3.1.

##### **4.4.2 The Search for Recipient Households**

At the first stage of the recipient search, with exact matching on age, sex, household structure and household size, 7,415 suitable households were found for the 12,719 donors. This number was increased to 10,585 after relaxing the matching criteria. This involved accepting as recipients households with the same number of adults (of the same sex and within one age group of those in the donor’s household), but not necessarily the same number, age or sex of children. In order to find recipients for the remaining 2,134 donors it was decided to relax the matching criteria further by removing the condition that recipients had to match the imputed individual on household structure. This left four donors for whom no suitable recipient could be found. Since Steps 2 and 3 together lead to over-imputation of individuals, these four individuals are excluded from the post imputation file.

#### **4.5 Pruning and Grafting**

After Steps 2 and 3, there were a total of 448,634 individuals in the post imputation file. Therefore, 3,367 imputed individuals need to be deleted to maintain consistency with the ‘true’ file. At this stage the program is undergoing testing. It is expected that empirical results will be presented at the Steering Committee Meeting.

#### **4.6 Imputation Results**

A full set of results based on the first data set will be presented at the Steering Committee Meeting.

#### **5.0 Discussion of the Results**

A discussion of the results so far will be presented at the Steering Committee Meeting. This will include an assessment of the initial empirical results together with suggested improvements and refinements for subsequent data sets.

#### **References**

Brown, J., Diamond, I., Chambers, R. and Buckner, L. Statistical Models to Estimate Underenumeration in the 2001 Census of England and Wales. Unpublished paper presented at the Population Association of America Annual Conference, Chicago, 2-4 April 1998.

## Appendix I – Illustrative Example of Iterative Scaling

In this simple example assume that the only LAD control totals for households are household size and tenure. The following table gives the true household distribution with totals of individuals in brackets.

		Size Class						
		1	2	3	4	5	6	
RENTS		400	300	100	100	50	50	1000 (2250)
OWNS		150	350	500	400	50	50	1500 (4500)
		550 (550)	650 (1300)	600 (1800)	500 (2000)	100 (500)	100 (600)	2500 (6750)

In general, we will only have ONC estimates of the marginal totals and not the individual cell totals. The next table gives the household distribution observed in the census. The uncalibrated household weights and their associated uncalibrated weighted marginal totals are in brackets.

		Size Class						
		1	2	3	4	5	6	
RENTS		270 (1.5)	250 (1.1)	100 (1)	100 (1)	40 (1.2)	40 (1.2)	800 (976)
OWNS		100 (1.5)	330 (1.05)	490 (1.05)	390 (1.05)	45 (1.1)	45 (1.1)	1400 (1519.5)
		370 (555)	580 (621.5)	590 (614.5)	490 (509.5)	85 (97.5)	85 (97.5)	2200 (2495.5)

These uncalibrated totals are not correct when compared to the ONC marginal totals for households by tenure or size. The first step of the iterative scaling algorithm to achieve this scales the weights to give the following weighted counts that match the marginal totals for tenure.

		Size Class						
		1	2	3	4	5	6	
RENTS		415	282	102	103	49	49	1000
OWNS		148	342	508	404	49	49	1500
		563	624	610	507	98	98	2500

The marginal distribution is now correct for tenure. The new weights are now scaled to match the household size marginal distribution. As the tenure distribution is now wrong this scaling is repeated until convergence is achieved. In this example after seven iterations the following table is obtained.

		Size Class						
		1	2	3	4	5	6	
RENTS		405	293	101	101	50	50	1000
OWNS		145	357	499	399	50	50	1500
		550	650	600	500	100	100	2500

This weighted table is now calibrated to both the marginal distributions.

## Appendix II – The CCS Design for the LAD Used

HtC Index	Number of EDs	ED Sample Size
Very Easy	144	14
Easy	210	16
Medium	186	19
Hard	193	18
Very Hard	197	18
TOTAL	930	85

The above table shows the distribution of enumeration districts by HtC index for the LAD used in the simulation study. The final column has the number of EDs sampled from each category of the index for each CCS generated in the simulation study.

### Appendix III – Additional Variables Used in the Coverage Modelling

Below is set out the additional variables that were calculated. The first is a household ethnicity variable that looks at the ethnicity of individuals within the household and puts the households into broad groups. This was used in model (1).

HHETHNIC	1	White
(Household	2	Black (Caribbean, African, Other)
Ethnicity)	3	Asian (Indian, Pakistani, Bangladeshi)
	4	Chinese / Other
	5	Mixed

For the individuals simplified variables were used for several variables. These are explained in the table below.

Complete Variable	Reduce Variable	Description
TENURE 1, 2 3 4, 5, 6 7, 8	TEN 1 2 3 4	Owner With Job, Farm, Shop Social Housing Private Landlord
ALWPRIM 1, 2, 3, 4 5 6 7 8 9 10, 11, 12	ECON 1 2 3 4 5 6 7	Working Government Employment Or Training Scheme Waiting To Take / Start A Job Unemployed / Looking For Work School / Full-Time Education Unable To Work (illness or disability) Economically Inactive
MARCON 1 2, 3 4, 5	MAR 1 2 3	Single (Never Married) Married Single (Previously Married)

## **Presented at ONC Steering Committee meeting on 13 November 1998**

### **Executive Summary of Imputation Results**

#### **Achievements:**

1. The full donor imputation system has been implemented on the first simulated dataset.
2. Results show that imputation works and can be used to provide a full One Number Census database.
3. For household variables at the LAD level, distributions after imputation are not significantly different from the true distributions.
4. For individual variables at the LAD level, while significantly different from the true distributions, the distributions after imputation are an improvement on the census.
5. In the post imputation database the bias across enumeration districts of census based estimates has been reduced with no increase in overall error.

#### **Lessons Learnt:**

1. Pruning and particularly grafting are extremely difficult to control and need to be minimised.
2. Imputation of completely missed households is relatively straightforward and computationally efficient.
3. In the search for recipient households in the imputation of individuals into counted households, it is computationally intensive, and not always possible, to find exact matches on age sex structure as well as household characteristics.
4. In current form, the process can be carried out for one large LAD using a reasonable PC in one week. With refinements it is expected that this time can be considerably reduced.

#### **Future Work:**

1. Investigate an intermediate calibration of individuals' weights to help minimise the need to prune and graft.
2. Investigate alternative definitions of impute classes

This will be done using the other nine datasets that have already been created. It should be noted that these datasets were created assuming a national CCS of 40,000 postcodes and it may be necessary to create new datasets under different assumptions.

## Pruning and Grafting

After Steps 2 and 3, there were a total of 448,634 individuals in the post imputation file. Therefore, 3,367 imputed individuals need to be deleted to maintain consistency with the ‘true’ file. The post-imputation (pre pruning and grafting) and ‘true’, target age-sex distributions are shown in Table 5. Table 6 shows the post-imputation and target household size distributions.

*Table 5: Distribution of Individuals by Age and Sex*

Age-sex group	True distribution (after pruning and grafting)	Imputed distribution (before pruning and grafting)	% relative difference <sup>1</sup>
Males 0-4	17777	17835	0.33
Males 5-9	17223	17411	1.09
Males 10-14	16588	16881	1.77
Males 15-19	15828	15672	-0.99
Males 20-24	16364	16231	-0.81
<b>Males 25-29</b>	<b>16637</b>	<b>15528</b>	<b>-6.67</b>
Males 30-34	15859	15426	-2.73
Males 35-39	15037	14926	-0.74
Males 40-49	15460	15543	0.54
Males 54-79	65340	66365	1.57
Males 80-84	2565	2489	-2.96
Males 85+	1172	1228	4.78
Females 0-4	16954	16923	-0.18
Females 5-9	16600	16759	0.96
Females 10-14	15442	15704	1.70
Females 15-19	15275	15544	1.76
<b>Females 20-24</b>	<b>17541</b>	<b>19121</b>	<b>9.01</b>
Females 25-29	17689	17860	0.97
Females 30-34	16655	16907	1.51
Females 35-39	14663	14845	1.24
Females 40-49	15461	15659	1.28
Females 54-79	74117	75282	1.57
Females 80-84	5490	5286	-3.72
Females 85+	3530	3209	-9.09
<b>TOTAL</b>	<b>445267</b>	<b>448634</b>	<b>0.76</b>

<sup>1</sup> % relative difference calculated as (imputed count-true count)/true count H 100

Table 6: Distribution of Households by Size

Household size	True distribution (after pruning and grafting)	Imputed distribution (before pruning and grafting)	% relative difference <sup>1</sup>
1	46033	44732	-2.83
2	53974	54520	1.01
3	27314	27195	-0.44
4	26209	26684	1.81
5	9790	10059	2.75
6	4122	4206	2.04
7	1619	1727	6.67
8	1003	964	-3.89
9	546	545	-0.18
10	300	292	-2.67
11	154	142	-7.79
12	77	73	-5.19
13	31	31	0.00
14	11	15	36.36
15	9	8	-11.11
16	8	4	-50.00
17	1	4	300.00
18	1	1	0.00
19	1	0	-100.00
20	2	2	0.00
23	1	1	0.00
TOTAL	171206	171205	0.00

<sup>1</sup> % relative difference calculated as (imputed count-true count)/true count H 100

## Imputation Results

Table 7: Distribution of Households by Tenure

Tenure Categories	True Distribution	Imputed Distribution	% relative difference <sup>1</sup>	Census Distribution
Owner occupied – mortgage	78,421	78,617	0.25	76,542
Owner occupied – outright	43,968	43,857	-0.25	43,108
With job, farm, shop	1,805	1,800	-0.28	1,749
Local authority	27,848	27,769	-0.28	26,709
New town corporation	416	416	0.00	392
Housing association	6,039	6,027	-0.20	5,715
Private landlord – furnished	5,689	5,737	0.84	4,667
Private landlord – unfurnished	7,020	6,983	-0.53	6,306
<b>TOTAL</b>	<b>171,206</b>	<b>171,206</b>	<b>0.00</b>	<b>165,188</b>

<sup>1</sup> % relative difference calculated as (imputed count-true count)/true count H 100

**NOTE:** The figures for the imputed distribution differ from those presented in Table 3 of the Steering Committee Paper. This is because at the final stage of pruning and grafting the number of households of size one is correct but the individuals are not quite correct by age and sex. This required the pruning off and then grafting on of approximately 800 single person households where there is little or no control on characteristics other than age and sex.

Table 8: Evaluation of Household Imputation

Variable	Truth	Imputed	% rel. dif. <sup>1</sup>	Census
<b>Building type</b>				
Caravan	282	276	-2.13	268
Detached	17906	17952	0.26	17640
Semi-detached	61878	62033	0.25	60657
Terraced (including end)	65195	65169	-0.04	62542
Purpose built flat (commercial block)	1956	1956	0.00	1834
Purpose built flat (in block of flats)	19771	19688	-0.42	18705
Converted house/flat (own entrance)	1125	1134	0.80	1050
Converted house/flat (shared entrance)	3093	2998	-3.07	2492
Chi-squared Value: tests against true distribution (p-value)		3.98 (0.782)		116.11 (0.000)
<b>Multi-occupied identifier</b>				
Not multi-occupied	166988	167074	0.05	161646
First household (multi-occupied)	1281	1266	-1.17	1107
Subsequent households (multi-occupied)	1798	1740	-3.23	1452
Last household (multi-occupied)	1139	1126	-1.14	983
Chi-squared Value: tests against true distribution (p-value)		2.24 (0.524)		73.52 (0.000)
<b>Type of accommodation</b>				
Not applicable	166988	167074	0.05	161646
One roomed flat	268	248	-7.46	208
One room or bedsit (not self-contained)	1178	1127	-4.33	875
Self-contained with 2+ rooms	2446	2460	0.57	2204
Not self-contained with 2+ rooms	326	297	-8.90	255
Chi-squared Value: tests against true distribution (p-value)		6.40 (0.171)		93.41 (0.000)
<b>Number of cars</b>				
0 cars	69662	69265	-0.57	66037
1 car	70706	70967	0.37	68805
2 cars	26064	26165	0.39	25643
3+ cars	4774	4809	0.73	4703
Chi-squared Value: tests against true distribution (p-value)		3.87 (0.275)		37.38 (0.000)

<sup>1</sup> % rel. dif. calculated as (imputed count-true count)/true count H 100

**NOTE:** The figures for the imputed distribution differ from those presented in Table 4 of the Steering Committee Paper. This is because at the final stage of pruning and grafting the household distributions were slightly altered as explained with Table 7.

Table 9: Evaluation of Individual Imputation

Variable	Truth	Imputed	% rel. dif. <sup>1</sup>	Census
<b>Ethnicity</b>				
White	374918	374287	-0.17	357890
Black Caribbean	3128	3190	1.98	2877
Black African	526	511	-2.85	473
Black Other	2189	2259	3.20	2009
Indian	11524	11539	0.13	10789
Pakistani	44518	44821	0.68	40929
Bangladeshi	3574	3648	2.07	3335
Chinese	646	639	-1.08	569
Other	4244	4373	3.04	3916
Chi-squared Value: tests against true distribution (p-value)		12.57 (0.128)		68.93 (0.000)
<b>Primary activity next week</b>				
Under 16	106509	106559	0.05	100651
Employee working full-time	126142	126984	0.67	120766
Employee working part-time	32122	32114	-0.02	31464
Self-employed, employing others	7153	7161	0.11	6944
Self employed, not employing others	13214	13325	0.84	12720
Govt employment/ training scheme	2655	2480	-6.59	2318
Waiting to take/start a job	591	551	-6.77	527
Unemployed/ looking for work	20741	19789	-4.59	17309
School/ full-time education	21508	21442	-0.31	19101
Unable to work (illness or disability)	13479	13558	0.59	13062
Retired from paid work	64393	64581	0.29	62638
Looking after home/family	36361	36321	-0.11	34917
Other economically inactive	399	402	0.75	370
Chi-squared Value: tests against true distribution (p-value)		65.81 (0.000)		486.12 (0.000)
<b>Tenure</b>				
Owner occupied – mortgage	247092	246074	-0.41	235709
Owner occupied – outright	93840	93721	-0.13	90669
With job, farm, shop	4695	4725	0.64	4481
Local authority (council)	61441	61808	0.60	58084
New town corporation	788	799	1.40	739
Housing association/ charitable trust	10581	10784	1.92	9902
Private landlord (furnished)	11717	12258	4.62	9706
Private landlord (unfurnished)	15113	15098	-0.10	13497
Chi-squared Value: tests against true distribution (p-value)		35.77 (0.000)		267.87 (0.000)

<sup>1</sup> % rel. dif. calculated as (imputed count-true count)/true count H 100

Table 9 shows that for variables at the individual level imputation has not been quite so successful. However, in all cases the distribution in the post imputation file is 'nearer' the truth than the census distribution.

To further investigate the results from the imputation the accuracy of estimates at the enumeration district has been considered. To do this two measures are calculated across enumeration districts for household and individual variables. The first is relative bias calculated as:

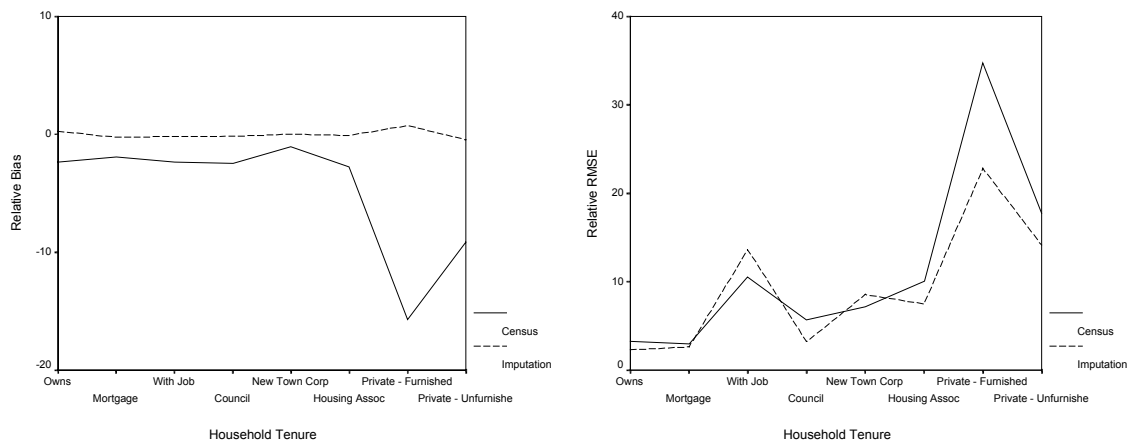
$$\text{Relative bias} = \frac{100}{\text{truth}} \times \frac{\sum_{e=1}^{930} (\text{estimate}_e - \text{truth}_e)}{930}$$

and the second is relative root mean square error (RMSE) calculated as:

$$\text{Relative RMSE} = \frac{100}{\text{truth}} \times \sqrt{\frac{\sum_{e=1}^{930} (\text{estimate}_e - \text{truth}_e)^2}{930}}$$

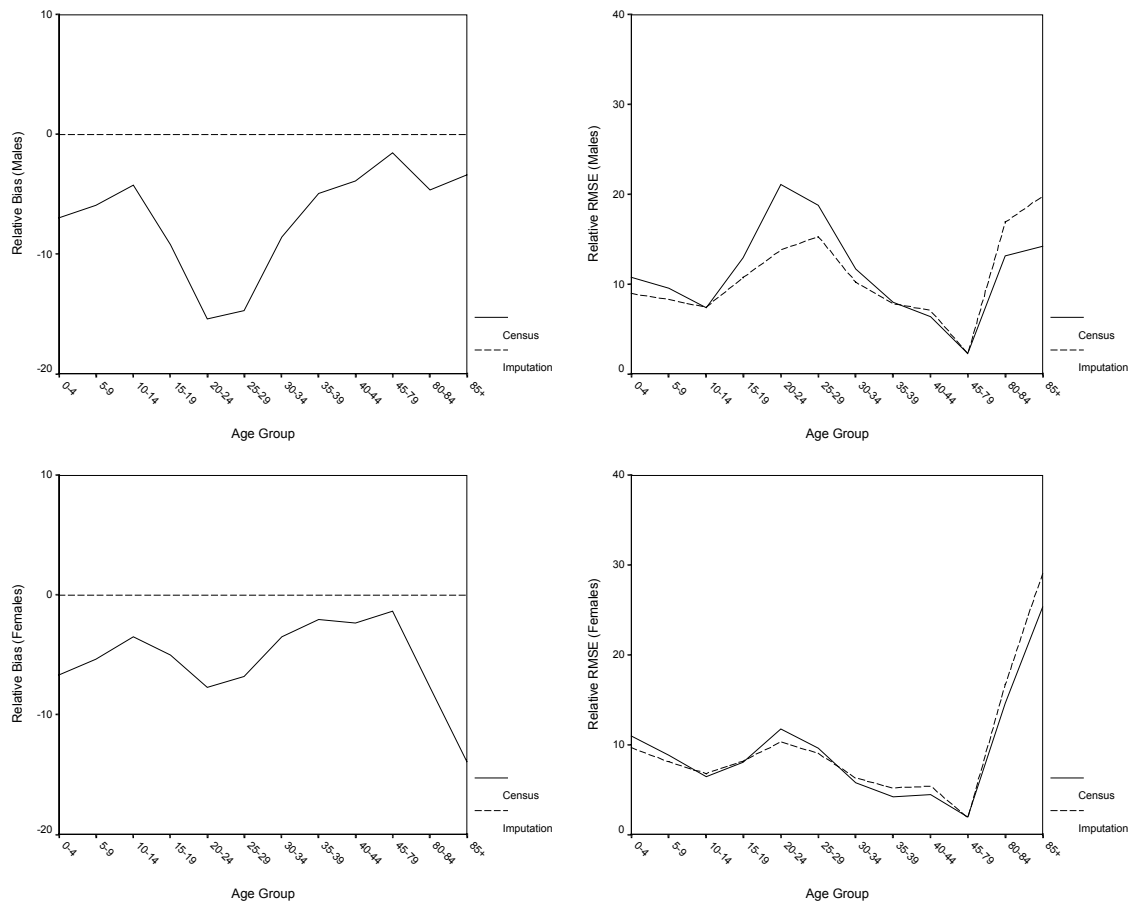
where e indexes each enumeration district. In both cases the estimate can be from the post imputation file or from the census file.

*Figure 4: Graphs to Compare the Census and Post Imputation Enumeration District Estimates for Tenure at the Household Level*



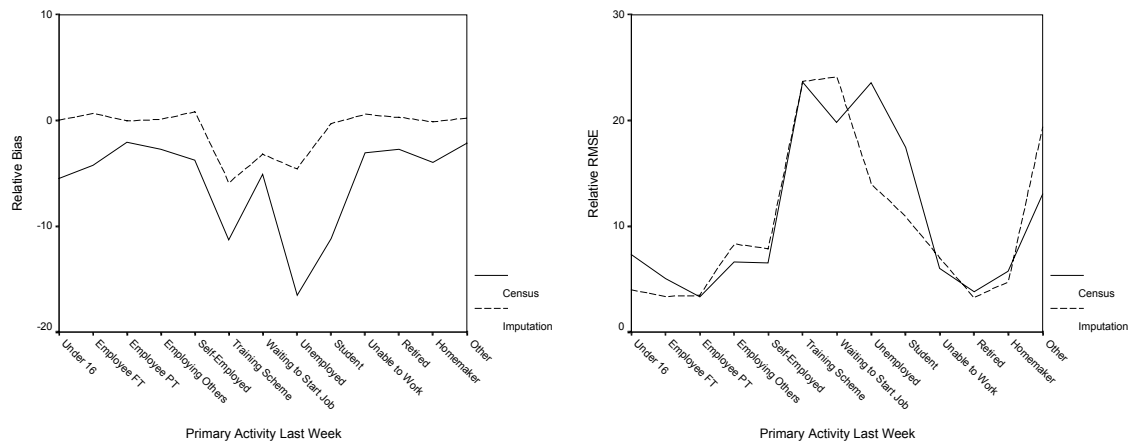
The relative bias graph in Figure 4 demonstrates the effect of the calibrating the household weights by tenure. The post-imputed estimates are basically unbiased across enumeration districts. The relative RMSE graph shows that this reduction in bias has been achieved at no increase to the overall error although the variance has increased.

Figure 5: Graphs to Compare the Census and Post Imputation Enumeration District Estimates for Age and Sex



Again the relative bias graphs in Figure 5 show that, in this case exact calibration from pruning and grafting, gives unbiased estimates across enumeration districts for age by sex after imputation. As before this has been achieved without increasing the overall error.

Figure 6: Graphs to Compare the Census and Post Imputation Enumeration District Estimates for Primary Activity Last Week



The relative bias graph in Figure 6 shows that when the weights have not been calibrated the post imputation estimates will not necessarily be unbiased. However, it does demonstrate that the uncalibrated weights still reduce significantly reduce the bias in the enumeration district estimates. As before, reduction in bias has been achieved with no real increase in overall error.