



ONS(ONC(SC))99/08

ONE NUMBER CENSUS STEERING COMMITTEE

A Donor Imputation System to Create a Census Database Fully Adjusted for Underenumeration

1. This paper was presented at the Statistics Canada Symposium, 2-5 May 1999 and is circulated for information.
2. **The Steering Committee are asked to note the paper.**

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

June 1999

Proceedings of the Statistics Canada Symposium 99
Combining Data from Different Sources
May 99

A DONOR IMPUTATION SYSTEM TO CREATE A CENSUS DATABASE FULLY ADJUSTED FOR UNDERENUMERATION

Fiona Steele¹, James Brown² and Ray Chambers²

ABSTRACT

Following the problems with estimating underenumeration in the 1991 Census of England and Wales the aim for the 2001 Census is to create a database that is fully adjusted for net underenumeration. To achieve this, the paper investigates weighted donor imputation methodology that utilises information from both the census and census coverage survey (CCS). The US Census Bureau has considered a similar approach for their 2000 Census (see Isaki *et al* 1998).

The proposed procedure distinguishes between individuals who are not counted by the census because their household is missed and those who are missed in counted households. Census data is linked to data from the CCS. Multinomial logistic regression is used to estimate the probabilities that households are missed by the census and the probabilities that individuals are missed in counted households. Household and individual coverage weights are constructed from the estimated probabilities and these feed into the donor imputation procedure.

The first stage of the imputation procedure uses household coverage weights in conjunction with geographical information to determine the number of households that need to be imputed in the local administrative area. Donor households are selected from counted households with characteristics similar to those of the missed households. The second stage of the procedure imputes individuals missed in counted households, based on individual coverage weights. A donor is selected for each imputed individual who is then placed in a suitable counted household. Finally, certain marginal totals generated from the post-imputation database need to be calibrated to already agreed estimates at the local administrative area. To achieve this for age-sex and household size, some addition and deletion of imputed individuals is necessary.

KEY WORDS: Controlled Imputation, Census Underenumeration, Weighting, Calibration

1. INTRODUCTION

The basic level of underenumeration for the national population estimated by a Post Enumeration Survey (PES) is a useful guide to how well the census has performed and adjustments at the regional level are often used in the re-basing of mid-year population estimates. However, national and local governments use the population census (usually every five or ten years) as the basis for planning and resource allocation down to very small areas. If underenumeration is uniform by geography and by the characteristics of individuals and households then it can effectively be ignored. However, this is generally accepted to not be the case. This paper describes an approach to this problem that adjusts the whole census database to reflect the level of underenumeration recorded at the national level. A similar approach has been investigated for adjustment of the US Census, see Isaki *et al* (1998).

The United Kingdom (UK) has faced the problem of underenumeration for a number of censuses with net underenumeration measured by a PES. In 1991, though, the PES was unable to estimate either the level of net underenumeration at the national level or allocate underenumeration to a regional level. To ensure that this does not occur in 2001 a major research project has been undertaken so that, in 2001, the Office for National Statistics (ONS) will be in a position to **estimate** and **adjust**

¹ Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, UK.

² Department of Social Statistics, University of Southampton, Southampton, Hants SO17 1BJ, UK.

accurately census outputs for net underenumeration. The overall ONC research is described in Brown *et al* (1999a). The first stage involves the estimation of underenumeration at sub-national administrative areas called Local Authority Districts or LADs. To achieve this, individuals on the census database in the PES sample areas must be ‘matched’ to the database generated by the PES, to be called the Census Coverage Survey (CCS) in the 2001 Censuses of the UK. This paper assumes that stage one has taken place and population estimates for these areas are available for a variety of basic demographic characteristics for both households and individuals. In what follows a controlled imputation methodology for stage two is presented. The aim is to adjust the individual level database so that it is consistent with the population estimates already produced at stage one. This is achieved using an imputation system driven by coverage weights estimated for households and individuals, which is described in Section 2. Section 3 presents a simulation study that investigates the performance of the controlled imputation methodology assuming population estimation from the CCS at stage one is achieved without error.

2. CONTROLLED IMPUTATION METHODOLOGY

There are a series of steps in the creation of a database that is fully adjusted for underenumeration.

- 1) Modelling the census coverage of households and individuals.
- 2) Imputation of households completely missed by the census.
- 3) Imputation of individuals missed by the census in counted households.
- 4) Final adjustments to the database in order to satisfy the consistency requirements for a ONC.

The methodology for each step is outlined in the following sections.

2.1 Step 1: Estimation of household and individual coverage weights

2.1.1 Derivation of household coverage weights

Following the census and the CCS each household within a CCS area can be placed in one of following four 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 the census and the CCS. While an unrealistic assumption, the households missed by both are accounted for in the dual system estimates at the design area³ level and the final imputed database is constrained to satisfy the totals at the CCS design area level. Excluding category 4, categories 1, 2 and 3 define a multinomial outcome that can be modelled for each LAD as follows:

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

where $\theta_{jke}^{(t)}$ is the probability that household j in postcode k in enumeration district (ED) e in an LAD with characteristics defined by Z_{jke} 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.

The estimated model for the CCS areas is extrapolated to non-CCS areas within the LAD to obtain predicted probabilities of being in a particular response category for each household. The probabilities for each response category estimated under model

³ The design area is a group of LADs for which the CCS can make direct estimates, see Brown *et al* (1999a, 1999b).

(1) are then used to calculate a coverage weight for each household (h/h) counted in the census that can be applied to the household database. The household coverage weight is defined as

$$w_{jke}^{h/h} = \frac{1}{\theta_{jke}^{(1)} + \theta_{jke}^{(3)}}$$

However, the resulting weighted sums of counted households will not, in general, match corresponding totals estimated for the LAD. Therefore the weights are calibrated to the LAD marginal totals for key household variables, such as tenure, using iterative proportional scaling.

2.1.2 Derivation of individual coverage weights

To calculate coverage weights for those individuals counted in counted households, two assumptions are necessary regarding coverage of individuals in CCS areas. If a household is only counted by the census, then no individuals from that household are missed by the census. Similarly, if only the CCS counts the household 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:

- a) Counted in the census, but missed by the CCS
- b) Counted in the CCS, but missed by the census
- c) Counted in both the census and the CCS

These categories are then used to define the outcome in another multinomial model:

$$\log \left(\frac{\pi_{ijke}^{(r)}}{\pi_{ijke}^{(c)}} \right) = \beta^{(r)} X_{ijke} + \gamma^{(r)} Z_{jke} \quad r = a, b \quad (2)$$

where $\pi_{ijke}^{(r)}$ is the estimated probability that individual i in household j in postcode k in ED e within an LAD with individual characteristics defined by X_{ijke} and household characteristics defined by Z_{jke} is in category r . (Model (2) uses category c 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 extrapolated to non-CCS areas to obtain predicted probabilities of being in a particular response category for each individual. The probabilities estimated under the model are used to calculate a coverage weight for each individual (ind) that can be applied to the individual database. The individual coverage weights are calculated as

$$w_{ijke}^{ind} = \frac{1}{\pi_{ijke}^{(a)} + \pi_{ijke}^{(c)}}$$

As before the resultant weighted sums of census counted individuals will not be equal to the corresponding LAD totals. At the final stage of the imputation procedure, further adjustments are necessary to meet agreed LAD totals by age, sex and household size. To minimise the amount of adjustment required at this stage, individual coverage weights are calibrated to the agreed age-sex totals following the household imputation but before the imputation of individuals.

2.2 Step 2: Imputation of households

The household-based file of counted households in an LAD is matched to the file of calibrated household coverage weights (as described in Section 2.1.1). This file is sorted by coverage weight, and by geographical location. For more efficient processing, households are then grouped into impute classes defined by the characteristics on which the household coverage weights are based. Weights are grouped into bands to give impute classes. The processing block is an impute class within an LAD.

Within each processing block, households are processed sequentially and running totals are retained 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 the weighted and unweighted running totals is less than or equal to 0.5. An imputed household is assigned a household coverage weight of zero.

In order to assign characteristics to the imputed households, a donor imputation method is used. For each 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.

A further source of information available in the UK is the 'dummy form'. This is a form completed by an enumerator which indicates the presence of a non-vacant household that has not been enumerated in the census. Previous censuses have shown that the quality of information collected in these forms is variable. However, the possibility of using them in the choice of location for imputed households is being investigated. This will require the capture of the information onto a computer database, something that has not been done in the past.

2.3 Step 3: Imputation of individuals into counted households

The individual weights estimated in Section 2.1.2 are not calibrated to population totals when calculated. However, it is necessary to do this to ensure that enough extra individuals with the correct characteristics are added. This is achieved by using iterative scaling to calibrate the weights to population totals that reflect the individuals already imputed by the household imputation described in Section 2.2.

The individual-based file of counted individuals is then sorted by weight, and by geographical location. Impute classes are defined by the characteristics on which the individual coverage weights are based. Individual coverage weights are grouped into bands to give impute classes. 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.

Individual and household characteristics are assigned to the imputed individuals in two separate stages. 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. The search for a donor is carried out in the same way as described above for the household imputation. The donor is selected at random from among the counted individuals with the same coverage weight and in the same ED as the counted individual that was processed immediately before the imputation. When a donor is found, 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.

In order to maintain sensible household structures for households into which individuals have been imputed, the type of recipient household sought depends on certain characteristics of the donor. In the simulation study that follows the choice of recipient depends on the age, marital status and household structure of the donor. Household structure is defined using both census and CCS information. Therefore, if an individual who was missed by the census is found in the CCS, the structure of their household will be edited accordingly. To illustrate the recipient search, consider an individual that the coverage weights suggest needs to be imputed. Suppose that a married person went missing from a 'couple without children' household. The household structure(s) that would result after exclusion of the imputed person defines the structure required for the recipient household. Thus the recipient for this individual must be a single person household. In this case, the marital status of the single person would be edited to married after the imputed person is added to the household. In a further attempt to maintain sensible households, the age-sex composition of the donor's household is also taken into account in the search for a recipient. After selection of a suitable recipient, the imputed individual is placed in the chosen household and is assigned the recipient's household characteristics.

2.4 Step 4: Final calibration ('pruning and grafting')

Due to the calibration of household coverage weights carried out before the household imputation, the number of households in each impute class will be within one household of the weighted total for that class. Further, the distribution of the household variables to which household weights are calibrated will be almost exactly the same as the target distributions. However,

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. In the final calibration stage, the post-imputation database is adjusted 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. To achieve this aim some addition and/or deletion of imputed individuals from imputed and counted households will be necessary.

The basic idea of the ‘pruning and grafting’ procedure is to start at the largest households and work down to households of size one, adding (‘grafting’) and deleting (‘pruning’) people to move households up or down in size. The addition of individuals follows the same process as individual imputation while the deletion is at random from a set of possible imputed individuals. This is controlled so that the age-sex distribution after pruning and grafting is exactly calibrated to the control distribution.

3 SIMULATION STUDY

3.1 Generation of the census and CCS data

As with any simulation study the exact nature of the results will depend on the way in which the data have been simulated. For this study ten censuses have been generated from a LAD of 1991 Census records using the same methodology applied to simulating censuses and CCSs in Brown *et al* (1999a and 1999b). The simulation population and CCS design is also the same as the population used for the simulations in Brown *et al* (1999b).

3.2 Generation of the households and individual coverage weights

For the simulated data set three multinomial models, using main effects only, have been estimated: one for household coverage, based on model (1) in Section 2.1.1, and two separate models for individual coverage of adults and children in counted households, both based on model (2) in Section 2.2.2. The explanatory variables used in the household model are tenure, household ethnicity, household structure, and the enumeration district’s HtC index. In the model for individual coverage within counted households children have been considered separately from adults, as they do not have an economic status (as measured by the census). The explanatory variables 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 all variables are based on the joint census-CCS data. If there is a conflict the census value is chosen unless it is due to an individual being missed.

The household coverage weights are calibrated 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 methodology rather than the ability to estimate totals at a higher level. The weights have been calibrated to the true distributions 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 carried out using an iterative scaling algorithm that converged very rapidly. The individual coverage weights are approximately (see Section 2.3) calibrated to marginal distributions after accounting for the individuals added by the household imputation. As with the household calibration the ‘true’ marginal distributions have been used.

3.3 Evaluation of the imputation procedure

The methodology described in Section 2 has been applied to the simulated census database and its associated household and individual coverage weights. The computer time taken to run the whole procedure is approximately 48 hours on a 450MHz Pentium II PC with 512 megabytes of RAM.

To evaluate the performance of the imputation methodology on the simulated census database, the marginal distributions of key household and individual characteristics in the unadjusted census and fully adjusted census databases are compared with their true distributions. Standard Pearson chi-square tests are used to test the hypothesis that the distribution of a variable in the adjusted census database is the same as its distribution in the true database. For a categorical variable with C classes the chi-square statistics is calculated as:

$$X^2 = \sum_{c=1}^C \frac{(T_c^{(adj)} - T_c)^2}{T_c}$$

where $T_c^{(adj)}$ is the number of households (or individuals) in class c in the fully adjusted census file, and T_c is the true number of households (individuals) in class c . The test statistic is compared to a chi-square distribution on $C-1$ degrees of freedom.

Chi-square tests are also used to compare distributions of household and individual variables in the unadjusted census database to their true distributions. In the calculation of the chi-square statistic to compare census counts with the truth, the true counts T_c are replaced by the census counts that would be expected if the census had the same percentage distribution as the true database. A comparison of these measures with those obtained for the adjusted census-truth contrast provides an indication of the improvement of the adjusted census database over the unadjusted census database.

Table 1: Evaluation of imputation procedure for selected household and individual variables

Household variable	X ² : tests against true distribution (p-value)		Individual variable	X ² : tests against true distribution (p-value)	
	Adj.	Unadj.		Adj.	Unadj.
<i>Tenure*</i>	0.73 (0.998)	179.94 (0.000)	<i>Ethnicity</i>	8.28 (0.407)	68.93 (0.000)
<i>Building type</i>	0.49 (0.999)	116.11 (0.000)	<i>Primary activity last week*</i>	6.43 (0.893)	486.12 (0.000)
<i>Number of cars</i>	0.43 (0.934)	37.38 (0.000)	<i>Tenure*</i>	3.44 (0.842)	267.87 (0.000)

*Note that coverage weights have been calibrated to ‘true’ LAD marginal totals for these variables.

Results from the chi-square tests for selected household and individual variables are presented in Table1. Before adjustment the marginal distribution of each of these variables differs significantly from the true distributions. However after controlled imputation has been applied to the census the marginal distributions at the LAD level are correct. This is expected for variables such as tenure that have been calibrated for both households and individuals, but not necessarily for variables such as ethnicity that are not calibrated.

While this is not a complete evaluation of the controlled imputation procedure, these initial results are extremely encouraging and demonstrate the feasibility of this methodology to create an adjusted census database. Future analysis will need to consider joint distributions at the LAD level and performance at the ED level.

REFERENCES

- Brown, J., Buckner, L., Diamond, I., Chambers, R. and Teague, A. (1999a) A Methodological Strategy for a One Number Census in the United Kingdom. To appear in RSS Series A, June 1999.
- Brown, J., Diamond, I., Chambers, R. and Buckner, L. (1999b) The Role of Dual System Estimation in the 2001 Census Coverage Surveys of the UK. Statistics Canada Symposium, Ottawa, 4th to 7th May, 1999.
- Isaki, C. T., Ikeda, M. M, Tsay, J. H and Fuller, W. A. (1998) A Transparent File for a One-Number Census. US Bureau of the Census 1998 Annual Research Conference, Washington DC, March 23-25 1998.