



ONS(ONC(SC))98/16

ONE NUMBER CENSUS STEERING COMMITTEE

One Number Census Methodology

1. This paper describes the proposed ONC methodology for 2001.
2. **The Steering Committee are asked to:**
 - a) **note the paper;**
 - b) **endorse the proposed methodology and**
 - c) **provide any comments at the meeting on the 13 November 1998, or in writing by 27 November 1998.**

**Ian Diamond
Department of Social Statistics
University of Southampton
Highfield
Southampton
Hampshire
SO17 1BJ**

October 1998

One Number Census Methodology

Ian Diamond

1. Background

One of the major uses of the decennial UK Census is in providing figures on which to rebase the annual population estimates. This base needs to take into account the level of underenumeration in the census, which has traditionally been measured from data collected in a post-enumeration survey (PES) and through comparison with the estimate of the population based on the previous census. Until the 1991 Census, there was close agreement between the adjusted census count (census + PES) and the estimate based on the previous census. Moreover, the estimated level of underenumeration was relatively small (less than one per cent). In 1991, the level of underenumeration was much higher (2.2 per cent); underenumeration did not occur uniformly across all socio-demographic groups and parts of the country (for example, it was estimated to be over 20 per cent for young males in inner cities); and there was a significant difference between the survey-based estimate and that rolled forward from the previous census. This necessitated the development of a deterministic approach to allocate adjustment.

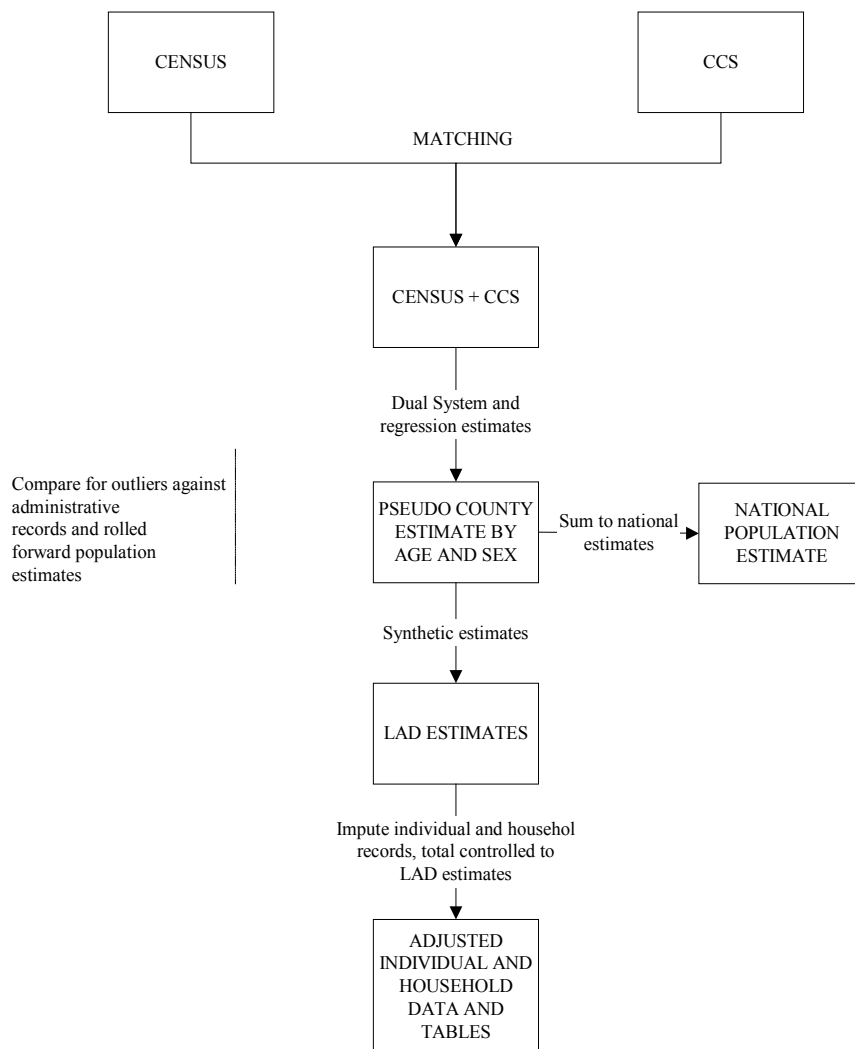
Despite efforts to maximise coverage in the 2001 Census, it is only realistic to expect there will be some degree of underenumeration. The One Number Census (ONC) project aims to measure this level of underenumeration in the most acceptable way, to provide a much clearer link than previously between the census counts and the population estimates, and to adjust all the census counts (which means the individual level database itself) for underenumeration. All counts will then add to 'One Number'. This has entailed the development of a new design of the post enumeration survey and how this should be integrated with other indicators of underenumeration provided by administrative records and demographic analyses.

The One Number Census process comprises six stages, which are illustrated in Figure 1. These include

1. A Census Coverage Survey (CCS) comprising a re-enumeration of a sample of postcodes. The survey will collect data on a small number of key variables central to measuring underenumeration together with a set of variables, which the user community believes are essential to have at a local level.
2. The CCS data will be matched, using a probability based matching procedure, against individual census records.
3. Combined regression and dual system estimation will be used to produce population estimates, by age and sex, for each area of a broad regional stratification of the UK. These regions are referred to 'design areas' in what follows. These are synonymous with pseudo counties which are large Local Authority District (LAD) or groups of smaller LADs.
4. LAD estimates will be derived from the Design Area Estimates using synthetic estimation.

5. National, design area and LAD estimates will be compared with a set of 1991 based estimates to assess their plausibility. In the event that any estimate is implausible a contingency strategy will be used.
6. Individual and household level records will be imputed for those individuals estimated to have been missed by the Census. Placement of these 'synthetic' records in the Census database will depend on the distribution of census coverage weights calculated from a multilevel analysis of the combined Census and CCS dataset. The number of records to be imputed will be constrained to each LAD estimate. Information on the distribution of dummy forms (identified non-responding households) in the census dataset will also be used in deciding where imputed records are to be placed in the database.

Figure 1: A Schematic overview of the One Number Census Process



2. The Design of the Census Coverage Survey

The aim of the CCS following the 2001 Census is to facilitate the estimation of underenumeration by age and by sex for design areas and to allocate this underenumeration down to small areas. The population size of the design areas is discussed in ONS(ONC(SC))98/12 and is intended to be around 500,000 people. These design areas for the CCS will be LADs or groups of LADs (very few LADs have populations greater than 500,000).

Following the 1991 Census a Census Validation Survey (CVS) was carried out in England, Scotland, and Wales. That survey aimed to estimate net underenumeration and to validate the quality of census data (Heady *et al.*, 1994). The second of these aims required a complete reinterview of a sample of households that had previously been enumerated in the census. This requirement was costly, due to the time required to fill out the complete census form, resulting in a small sample size. It also meant that the ability of the CVS to find 'missed' households was compromised, since no independent listing of households in selected postcodes was carried out. The CCS in 2001 will address coverage exclusively, with selected postcodes independently relisted in an attempt to identify all households in the area. In addition, focussing on coverage allows for a much shorter doorstep questionnaire. Savings in time can be translated into a larger sample size. Information on the quality of census data will be obtained from other sources, particularly the question testing programme, the 1997 Census Test and through a separate quality survey carried out in 1999.

The CCS will be a postcode-unit based survey. This requires the re-enumeration of a sample of postcode units rather than households. Although it is technically feasible to design a household-based CCS by sampling delivery points on the UK Postal Address File (PAF), lack of complete coverage by this sample frame make it unsuitable for checking coverage in the Census. Consequently an area-based sampling design has been chosen for the CCS, with census enumeration districts (EDs) as primary sampling units (PSUs) and postcodes within EDs as secondary sampling units (SSUs). Sub-sampling of households within postcodes was not considered since coverage data from all households in a sampled postcode is necessary for estimation of small area effects in the multilevel models proposed for stage six of the ONC.

Subject to resource constraints, the CCS sample design will be optimised to produce population estimates of maximal accuracy for the 24 age-sex groups defined by sex (male/female) and 12 age classes: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-79, 80-84, 85+. The ages 45-79 have been combined since there was no evidence of any marked underenumeration in this group in 1991. This age grouping will be reviewed prior to finalising the CCS design.

It is expected that underenumeration in the 2001 Census will be higher in areas characterised by particular social, economic and demographic characteristics. For example, people in dwellings occupied by more than one household (multi-occupancy) have a relatively high probability of not being enumerated in a census. In order to control for this effect EDs within each design area are stratified by a 'Hard to Count' (HtC) score. This score was chosen to represent social, economic and demographic characteristics that were found to be important determinants of underenumeration by the Office for National Statistics (ONS) and the Estimating With Confidence Project (Simpson *et al.*, 1997). The variables making up the HtC score will be reviewed prior to finalisation of the CCS design. The 'prototype' HtC score used in CCS development has been based on the following variables from the 1991 Census:

- percentage of heads of household who experienced language difficulty as defined by country of birth;
- percentage of young people who migrated into the enumeration district in the last year;
- percentage of imputed residents for the enumeration district;
- percentage of households in multiply-occupied buildings; and
- percentage of households which were privately rented.

For the purpose of sample design, the HtC scores have been converted to a five point HtC index by dividing them into quintiles at a national level, with each quintile assigned an index value from 1 (easiest to count) to 5 (hardest to count). The stratification used in the CCS design is then based on ED values of this HtC index as well ED size, as measured by projected population count at the 2001 Census. Note that preliminary evaluation studies of the CCS design have substituted actual 1991 Census counts for these projected counts.

All design areas in the CCS design are treated in the same way, the CCS design for a single design area is now described. Within such an area, a robust approach has been adopted for the design of the first stage of the CCS sample (selection of EDs). This assumes that within strata defined by values of the HtC index and by size ranges corresponding to ‘projected’ 2001 census counts, the true 2001 ED population counts for each of the 24 age-sex groups of interest will be independently and identically distributed. The actual allocation of the sample of EDs between these strata is then designed to minimise the sampling variability of a stratified expansion estimate of the design area total of a ‘design variable’ constructed as a linear combination of key age-sex counts for each ED. Stratification by the HtC index is important as the level of undercount will depend on the characteristics of the EDs. It also ensures that the CCS sample is spread across the full range of EDs. Further stratification by size based on projected 2001 census counts improves efficiency by reducing the within stratum variance of the design variable, and, by construction, the corresponding variances of all 24 age-sex counts. Ideally one would like to use the actual 2001 counts for this size measure, but the timing of the CCS makes this impossible.

The second stage of the CCS design consists of the random selection of a fixed number of postcodes within each selected enumeration district. Since this subsampling will result in a loss of efficiency, it is proposed that a regression estimator be used rather than the simple stratified expansion estimator underpinning the design discussed above. This estimator is discussed in Section 4.

3. Matching the Census and CCS Records

The estimation strategy described in section 4 requires that one can identify the number of individuals and households observed in both Census and CCS and those observed only once. Consequently census and CCS records must be matched. An underenumeration of around two to three percent nationally means that although the absolute numbers may be large, percentages are small. Thus the ONC process requires an accurate matching methodology.

The independent enumeration methodologies employed by the Census and CCS means that simple matching using a unique identifier common to both lists is not possible. Furthermore, simple exact matching on the variables collected in common by both methods is out of the question as there will be errors in both sets of data caused by incorrect recording, misunderstandings, the time gap, errors introduced during processing etc. The size of the CCS

also means that hand matching is not feasible. Thus a largely automated process involving probability matching is necessary.

Probability matching involves assigning a probability weight to a pair of records based on the level of agreement between the two records. The probability weights reflect the likelihood that the two records correspond to the same individual. First, blocking variables must be identified. A blocking variable, e.g. a postcode, is used to reduce the number of comparisons required by an initial grouping of the records. Probability matching is only undertaken within blocks as defined by the blocking variable.

Matching variables such as name, tenure and month of birth are then compared for each pair of records within a block. Provided the variables being compared are independent of each other, the probability weights associated with each variable can be summed to give an overall probability weight for the two records. Records are matched if the likelihood of them relating to the same individual exceeds an agreed threshold.

Before the probability weights can be calculated it is necessary to perform a clerical matching exercise to produce a file of matched records. This file should be representative of the total set of records to be matched. The file is then used to attain the likelihood that a given pair of outcomes are observed given that the records belong to the same person, for each possible pair of outcomes.

The CCS collects data for two purposes; to enable the data to be matched against the census; and to identify the characteristics of underenumeration via the modelling process, so that adjustments can be applied to the whole population. In order that the second part is not biased by the first the matching and modelling variables should be as independent as possible.

As the data are structured both geographically and by individuals within households we utilise this structure within the matching strategy.

The key stages of the matching are as follows:

1. Produce a clerically matched file
2. Derive the matching weights
3. Use 'blocking' variables to reduce the number of comparisons made
4. Match households
5. Match individuals within matched households
6. Clerically check any CCS forms left unmatched.

More details of the proposed matching methodology is given in ONS(ONC(SC))98/14.

4. Estimation of Design Area Age-Sex Populations

There are two stages of estimation in the CCS. First, a dual system estimation (DSE) method is used to estimate the number of people in different age-sex groups missed by both census and the CCS within each postcode in the CCS sample. Second, the postcode level population counts obtained from these DSEs are used in regression estimation to obtain final counts for the design area as a whole.

To start, we describe the DSE component of this methodology. This is based on the fact that it is unlikely that the union count (i.e. the total of those counted in the census and in the CCS) for an area will constitute a complete count. DSE assumes that (i) the census and CCS counts are independent and (ii) the probability of ‘capture’ by one or both of these counts is the same for all individuals in the area of interest. When these assumptions hold, DSE gives an unbiased estimate of the total population. Hogan (1993) describes the implementation of DSE for the 1990 US Census. In this case assumption (i) was approximated through the operational independence of the Census and PES data capture processes, and assumption (ii) was approximated by forming post strata based on characteristics believed to be related to heterogeneity in the capture probabilities.

In the context of the CCS, DSE will be used with the census and CCS data as a method of improving the population count for a sampled postcode, rather than as a method of estimation in itself. That is, given matched census and CCS data for a CCS postcode, DSE is used to define a new count which is the union count plus an adjustment for people missed by both the census and the CCS in that postcode. This ‘DSE count’ for the sampled postcode is then used as the ‘dependent’ variable in a regression model, which links this count with the census count for that postcode.

This regression model is based on the assumption that the 2001 Census count and the dual system adjusted CCS count within each postcode satisfy a linear regression relationship, with the possibility of a non-zero intercept as in some postcodes the census can miss all the people from a certain age-sex group. Given that it is known from the 1991 Census that undercount varies by age and sex as well as by local characteristics, a separate regression model within each age-sex group for each HtC category within each design area is used. Let Y_{id} denote the adjusted CCS adjusted count for a particular age-sex group in postcode i in HtC group d in a particular design area, with X_{id} denoting the corresponding 2001 Census count. Estimation in the CCS will be based on the simple regression model:

$$\left. \begin{aligned} E\{Y_{id}|X_{id}\} &= \alpha_d + \beta_d X_{id} \\ \text{Var}\{Y_{id}|X_{id}\} &= \sigma_d^2 \end{aligned} \right\} i \in d \quad (1)$$

$$\text{Cov}\{Y_{id}, Y_{jf} | X_{id}, X_{jf}\} = 0 \quad \text{for all } i \neq j$$

Substituting ordinary least squares (OLS) estimators for α_d and β_d into (1), it is straightforward to show (Royall, 1970) that under this model the Best Linear Unbiased Predictor (BLUP) for the total count T of the age-sex group in the design area is:

$$\hat{T} = \sum_{d=1}^5 \left\{ T_{Sd} + \sum_{i \in R_d} (\hat{\alpha}_d + \hat{\beta}_d X_{id}) \right\} = \sum_{d=1}^5 \hat{T}_d \quad (2)$$

where T_{Sd} is the total adjusted CCS count for the age-sex group for CCS sampled postcodes in category d of the HtC index in the design area; and R_d is the set of non-sampled postcodes in category d of the HtC index in the design area. Strictly speaking the model specified by (1) is known to be wrong on at least two counts. The first is the assumption of constant residual variation, which is unlikely unless the X_{id} are all approximately the same within the age-sex group in the design area. The other is that the zero covariance assumption in (1) ignores

correlation between postcode counts within a ED. However, the simple OLS estimator (2) remains unbiased under both types of mis-specification, and the OLS estimator is only marginally inefficient under a nonzero covariance structure (Scott and Holt, 1982).

The variance of $\hat{T} - T$, the estimation error associated with (2), can be estimated using the model (1). Unlike (2), this is sensitive to mis-specification of the variance structure (Royall and Cumberland, 1978). Consequently, as the postcodes are clustered within EDs, the conservative ultimate cluster variance estimator will be used. This is given by

$$\hat{V}(\hat{T} - T) = \sum_{d=1}^5 \frac{1}{m_d(m_d - 1)} \sum_{e=1}^{m_d} (\hat{T}_d^{(e)} - \hat{T}_d)^2 \quad (3)$$

where $\hat{T}_d^{(e)}$ denotes the BLUP for the population total of category d of the HtC index based only on the sample data from ED e .

The above estimation strategy represents a regression generalisation of the Horvitz-Thompson DSE estimator proposed in Alho (1994). As a postcode is a small population in a generally small geographic area, and with the counts split by age and sex, the DSE homogeneity assumption should not be seriously violated. In the situation where people missed by the census have a higher chance of being missed by the CCS than those counted by the census, one would still expect the regression estimator based on the DSE count to underestimate but to a lesser extent than the regression estimator based on the union count. When the reverse happens and the CCS is very good at finding the missed people (the requirement for getting unbiased estimates when using the union count in the regression estimator) one would expect the DSE count regression estimator to overestimate. However, unless these dependencies are extremely high one would not expect a gross error.

5. Local Authority District Estimation

Direct estimation using the CCS only produces estimates by age and sex for each design area in the UK. In the case of a LAD with a population of 500,000 or above this will give a direct estimate of the LAD population by age and sex. However, for the smaller LADs grouped to form design areas this will not be the case although all LADs will be sampled in the CCS. For these small LADs it will be necessary to carry out further estimation step, and allocate the design area estimate to the LADs constituting this area.

Standard small area synthetic estimation techniques are used for this purpose. These techniques are based on the idea that a statistical model fitted to data from a large area (in our case the CCS design area) can be applied to a much smaller area to produce a synthetic estimate for that area. The problem with this approach is that while the estimators based on the large area model have small variance they are usually biased for any particular small area. A compromise, introduced in the 1980s, involves the introduction random effects for the small areas into the large area model. These allow the estimates for each small area to vary around the synthetic estimates for those areas. This helps reduce the bias in the estimate for a small area at the cost of a slight increase in its variance (Gosh and Rao, 1994).

As described in the previous section, direct estimation at the CCS design area is based on the linear regression model (1) linking the 2001 Census count for each postcode with the DSE-

adjusted CCS count for the postcode. This model can be extended to allow for the multiple LADs within a CCS design area by writing it in the form

$$Y_{idl} = \alpha_d + \beta_d X_{idl} + \delta_{dl} + \varepsilon_{idl}$$

where the extra index $l = 1 \dots L$ denotes the LADs in the design area has been introduced, δ_{dl} represents an LAD ‘effect’ common to all postcodes with HtC index d , and ε_{idl} represents a postcode specific error term. The addition of the δ_{dl} term above represents differences between the LADs that have been grouped to form the design area.

This regression model can be fitted to the CCS data for a design area, and the LAD effects δ_{dl} estimated. For consistency, LAD population totals obtained in this way will be adjusted so that they sum to the original CCS design area totals, and they are always at least as large as the 2001 Census counts for the LAD.

6. Imputation of Missed Household and Individuals

6.1 Introduction

This 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 which are then calibrated so that weighted counts of individuals and households found by the Census agree with the corresponding CCS estimates for the total population by age-sex group and by household size for each LAD. These calibrated coverage weights are then used in a donor imputation system to create the missed households and individuals.

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.

6.2 Creating Household Coverage Weights

After the census and the CCS it can be assumed that all households within CCS areas fit 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.

Underlying this is the assumption that no household is missed by both. While this is an unrealistic assumption the households missed by both are accounted for by the CCS estimation

process and the final imputed database is constrained to satisfy these estimated totals at the CCS design area level. The categories (1) - (3) above define a multinomial outcome variable that can be modelled for each LAD using a logistic specification. Based on this model, the probability $\theta_{jidl}^{(t)}$ that household j in postcode i in HtC group d in LAD l has outcome t can be estimated. For outcomes $t = 1$ and $t = 3$ this estimated probability will be a function of the characteristics of the household as measured by the Census. This model can therefore be extrapolated to non-CCS areas to obtain estimated coverage probabilities for all households. Consequently, for each household j counted in the Census a household (h/h) coverage weight

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

can be calculated. In general, the weighted sums of households of different sizes computed using these weights will not agree with the corresponding CCS estimates for the LAD. Consequently, these weights are calibrated (via an iterative scaling procedure) so these constraints are satisfied.

6.3 Creating Individual Coverage Weights

Corresponding coverage weights for individuals counted by the Census are obtained using similar assumptions. In this case it is assumed that if a household is only counted by the census then no individuals from that household are missed by the census, and similarly, if the household is only counted by the CCS then no individuals from that household are missed by the CCS. Although this assumption is, in practice, violated the extra people are again accounted for by constraining to CCS estimated totals at the LAD level. Using these assumptions it is only necessary to consider individuals in household counted by both the Census and the CCS. In this case the possible categories are:

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

Again, matched Census/CCS data and an assumed multinomial logistic model are used to estimate the probability $\pi_{kjidl}^{(r)}$ that individual k in household j in postcode k in HtC group d in LAD l has outcome r . As with the household model the individual probabilities for outcomes $r = 11$ and $r = 13$ depend on individual and household characteristics as measured in the Census and so can be extended to allow computation of coverage probabilities for all individuals counted by the Census within households also counted by the Census. For each such individual (ind), therefore, a coverage weight

$$w_{kjidl}^{ind} = \frac{1}{\pi_{kjidl}^{(11)} + \pi_{kjidl}^{(13)}}$$

can be calculated.

6.4 Donor Imputation for Missed Households

The next stage in the process uses the household weights to impute completely missed households. In order to do this, households are split 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.

6.5 Donor Imputation for Missed Individuals

This is the most complex stage of the imputation due to the fact that adding individuals to households naturally changes the structure of the recipient household. This stage is best thought of 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.

6.6 Pruning and Grafting of Individuals

The two 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.

Further details on the imputation process are given in ONS(ONC(SC))98/15.

References

- Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics*, **10**, 245 - 256.
- Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: An appraisal. *Statistical Science*, **9**, 55-93.
- Heady, P., Smith, S. and Avery, V. (1994) *1991 Census Validation Survey: Coverage Report*, London: HMSO.
- Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *J.A.S.A.*, **88**, 1047-1060.
- Royall, R. M. (1970) On finite population sampling under certain linear regression models. *Biometrika*, **57**, 377-387.
- Royall, R. M. and Cumberland, W. G. (1978) Variance estimation in finite population sampling. *J.A.S.A.*, **73**, 351-361.
- Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *J.A.S.A.*, **77**, 848-854.
- Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends*, **90**, 31-39.