



ONS(ONC(SC))00/01

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 9th February 2000, or in writing by 23rd February 2000.**

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

January 2000

One Number Census Methodology

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, traditionally this has been measured from data collected in a post-enumeration survey (PES) and (at the national level) through comparison with the estimate of the population based on the previous census. In the 1991 Census, although the level of underenumeration was not high (estimated at 2.2 per cent), it did not occur uniformly across all socio-demographic groups and parts of the country. There was also a significant difference between the survey-based estimate and that rolled forward from the previous census. Further investigation showed that the PES had failed to measure the level of underenumeration and its degree of variability adequately.

Maximising coverage in the 2001 Census is a priority. A number of initiatives have been introduced to help achieve this, for example:

- the Census forms have been redesigned to make them easier to complete;
- population definitions for the Census have been reviewed;
- postback of Census forms will be allowed for the first time; and
- resources will be concentrated in areas where response rates are lowest.

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 underenumeration, provide a clear link between the Census counts and the population estimates, and adjust all Census counts (which means the individual level database itself) for underenumeration.

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

1. A Census Coverage Survey (CCS) will re-enumerate a sample of postcodes (geographical units of around 15 households). The survey will collect data on a small number of key variables central to measuring underenumeration.
2. The CCS data will be matched, using a probability based matching procedure, against individual Census records.
3. Combined ratio and dual system estimation will be used to produce estimates of the population based on the Census and CCS, by age and sex, for each area of a broad regional stratification of the UK. These regions, each with a population of around 0.5 million, are referred to as 'Design Groups' and are large Local Authority Districts (LADs) or groups of smaller LADs. The size of the Design Groups was selected to ensure a high efficiency of the design, based on a simulation study. LADs are important units of resource allocation in the UK. There are over 400 LADs of varying population sizes.
4. LAD estimates will be derived from the Design Group estimates using synthetic estimation.
5. National, Design Group 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 estimated to have been missed by the Census.

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

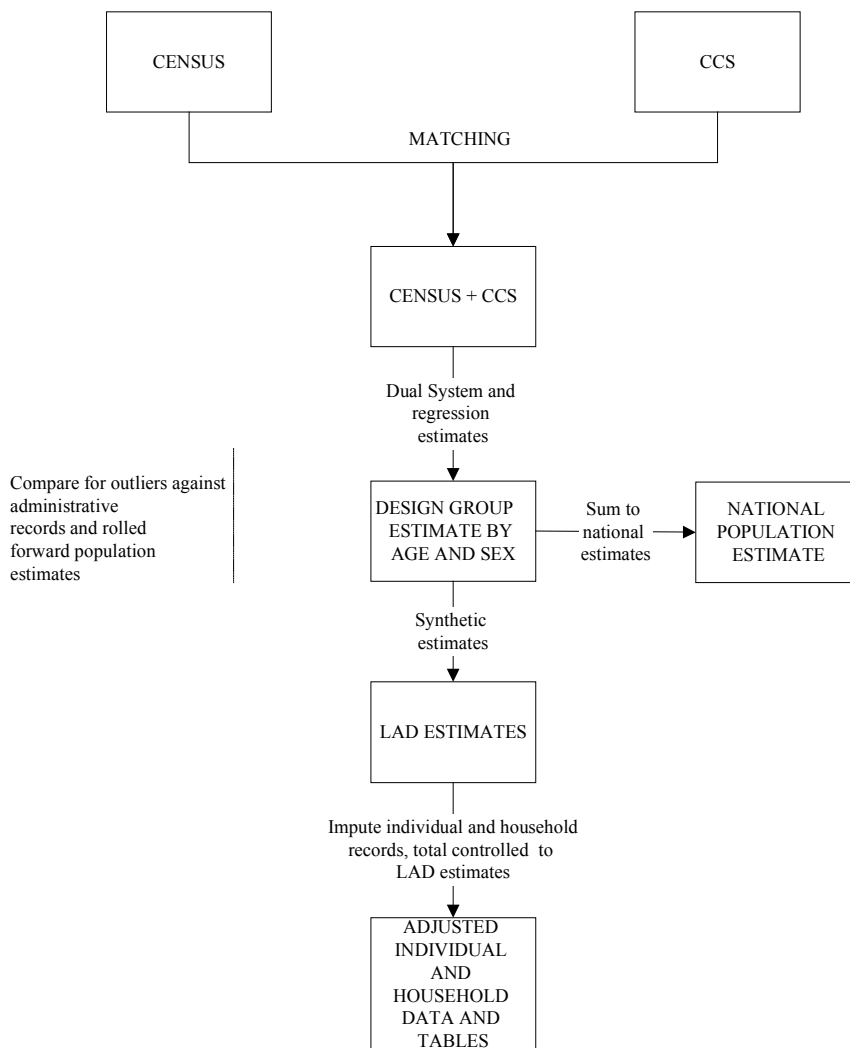


Figure 1 illustrates the One Number Census process, which comprises six stages:

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 all Local Authority Districts (LADs) in England and Wales. However, a CCS Design with the objective of producing direct estimates for Local Authority Districts would lead to a prohibitively large sample size. Therefore, to allow a more efficient sampling strategy, geographically contiguous LADs are aggregated to form Design Groups of population 500,000. These Design Groups are then used independently throughout the whole ONC process as strata for design, estimation and imputation. The population size of the Design Groups was investigated in ONS(ONC(SC))98/12 and the actual groupings are presented in ONS(ONC(SC))00/10.

As a result of this work, subject to resource constraints, the CCS sample design will be optimised to produce population estimates for the Design Groups of maximal accuracy for the 36 age-sex groups defined by sex (male/female) and 18 age classes: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+.

2.1 Sampling Units

The CCS will be a postcode-unit based survey, re-enumerating a sample of postcode units rather than households. It is technically feasible to design a household-based CCS by sampling delivery points on the UK Postal Address File (PAF), but the lack of complete coverage of this sample frame makes it unsuitable for checking coverage in the Census. Consequently, an area-based sampling design was chosen for the CCS with postcode units as the area. Stratifying variables at the postcode level beyond an estimate of the number of addresses are not known, and therefore postcodes are linked to 1991 Census Enumeration Districts (EDs) for which there is a wealth of reliable micro level data. The CCS will be a two stage cluster design with 1991 EDs as primary sampling units (PSUs) and postcodes within EDs as secondary sampling units (SSUs). The following section describes the stratification of the PSUs.

2.2 Stratification of PSUs

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 Group are firstly stratified by a national 'Hard to Count' (HtC) score. This score is chosen to represent some of the 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 current 'best guess' HtC score is an extended version of that used in the 1999 Rehearsal. It is based on the following variables from the 1991 Census:

- counts of households containing young people who migrated into the enumeration district in the last year;
- counts of imputed households for the enumeration district;
- counts of households in multiply-occupied buildings; and
- counts of households which were privately rented.

These variables making up the HtC score will be reviewed following the completion of the 1999 Census Rehearsal evaluation. The score is derived by dividing the sum of the variables by the total number of households in the ED. For the purpose of sample design, the HtC scores will be converted to a three point HtC index by dividing the EDs into a 40%, 40%, 20% distribution at a national level, with each group assigned an index value from 1 (easiest to count) to 3 (hardest to count). This approach will be reviewed following the 1999 Census Rehearsal evaluation. The stratification used in the CCS design is then based on ED values of this HtC index as well as ED size, as measured by population count at the 1991 Census. The 1991 Census counts are used as a proxy for the unknown 2001 ED population counts.

2.3 Overall Design

All Design Groups are treated in the same way as each other. Within a Design Group, a robust approach has been adopted for the design of the first stage of the CCS sample. This assumes that within the strata defined by the HtC index and by size ranges corresponding to 1991 census counts, the true 2001 ED population counts for each of the 36 age-sex groups of interest will be independently and identically distributed. The 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 Group total of a 'design variable'. This measure is constructed as a linear combination of key age-sex counts for each ED. The key age-sex groups used are those that experience the greatest underenumeration in past censuses, and hence are likely to be those with the greatest variability.

They are males aged 0-4, females aged 0-4, males aged 20-24, males aged 25-29, males aged 30-34 and females aged 85+.

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 1991 census counts improves efficiency by reducing the within stratum variance of the design variable, and, by construction, the corresponding variances of all 36 age-sex counts. Ideally the actual 2001 counts would be used for this size stratification, but the timing of the CCS makes this impossible. The selection of the primary sampling units will also ensure that each LAD in the Design Group is represented in the sample.

The second stage of the CCS design consists of the random selection of postcodes within each selected primary sampling unit. The number of postcodes to be chosen within each ED was investigated in paper ONS(ONC(SC))98/12. The research indicated that a maximum of five postcodes per ED would provide an efficient allocation of resources while still maintaining a robust approach.

Since this subsampling will result in a loss of efficiency, it is proposed that a ratio type estimator be used rather than the simple stratified expansion estimator underpinning the design discussed above. The estimator is described in Section 4.

2.4 Sample size

To achieve the aims of the CCS, the sample size must be sufficiently large to enable population estimates of an acceptable degree of precision. Through simulating the design described above, research indicated that the optimal sample size representing the best value for money in terms of precision is 20,000 postcodes for England and Wales. This research is presented in ONS(ONC(SC))98/12.

The sample will be allocated to each Design Group using a similar strategy to the first stage CCS design. England and Wales will be treated as a single Design Group, stratified by the HtC index. The design variable used in the sample size calculation for a single Design Group will be used to determine the national sample size within each HtC strata. These sample sizes will then be proportionally allocated to the Design Groups. This results in an allocation that is slightly weighted towards the harder to count areas.

2.5 Variation with GRO(S)

GRO(S) are adopting a similar design at the first stage, ie they will select enumeration districts in the same way. However two considerations predicate against the use of a fixed five postcodes per ED in Scotland. First, Scotland has a number of very small administrative districts particularly in the north of the country; and second, although the majority of the population live in urban areas there are large rural areas characterised often by small postcodes. On the first point, there is a need to have sample points in all the administrative districts and on the second, it is possible that in rural areas, five postcodes may, in total, have few households and that these may not constitute a workload for an interviewer. It is proposed, therefore, to adopt a 'workload' based scheme in Scotland whereby postcodes will be sampled until around 100 addresses have been counted. This strategy is currently being simulated to provide evidence on expected accuracy. However it is expected that the design will have largely the same degree of accuracy but that it will result potentially in an increase in the sample size in particular in rural areas. It should be noted that, as there are relatively few people in the rural areas this should not result in a great increase in the sample size.

2.6 History

This design evolved as follows:

- a) Initial research suggested that a PES would be essential. This was endorsed by the Steering Committee on 12 June 1997.
- b) The strategy of re-enumerating postcodes and using a stratified two stage sample, proposed on 12 June 1997 was approved in principle on 27 November 1997 because of the lack of a suitable framework as it permitted an independent assessment of all aspects of the enumeration.
- c) The efficiency of the design and decisions regarding sample size were approved on 13 November 1998.

Issues still to be resolved following evaluation of Dress Rehearsal include

- d) Composition of Hard to Count (HtC) index.
- e) Number of categories and distribution of HtC index.

Note that it may be necessary, because of time constraints to make judgements on these last two issues.

3. Matching the Census Coverage Survey and Census Records

The estimation strategy outlined in Section 4 requires the identification of the number of individuals and households observed in both the Census and CCS and those observed only once. Underenumeration of around two to three percent nationally means that, although 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 mean 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 entails assigning a probability weight to a pair of records based on the level of agreement between them. The probability weights reflect the likelihood that the two records correspond to the same individual. A blocking variable, e.g. 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 variables.

Matching variables such as name, type of accommodation and month of birth are 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, for the Census record that most closely resembles the CCS record in question, the likelihood of them relating to the same household or individual exceeds an agreed threshold.

The CCS data will be used 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.

The initial probability weights used in 2001 will have been calculated from the data collected during the 1999 Census Rehearsal. These weights will be refined as the 2001 matching process progresses. As the data are structured both geographically and by individuals within households this structure will be utilised within the matching strategy.

The key stages of the matching are as follows:

1. Use blocking variables to reduce the number of comparisons made
2. Match households
3. Match individuals within matched households
4. Clerically check any CCS forms left unmatched.

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

3.1 History

Matching was recognised as a key element of estimation process. The strategy was evolved over the winter of 1997/98 and included external consultants with a great deal of experience in matching records such as these. The initial strategy was endorsed on 13 November 1998. It was agreed that in order fully to finalise the methodology, data from the dress rehearsal would be needed.

4. Estimation of Design Group 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 accounting for individuals 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 ratio estimates to obtain final counts for the Design Group as a whole.

4.1 Dual System Estimation

DSE estimates the total population accounting for individuals missed by both the census and the CCS. It does this by assuming 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. The advantage of using the DSE at the postcode level, and controlling for age and sex, is that the assumptions of homogeneity and independence will be more closely met. However, simulations presented in ONS(ONC(SC))00/03A show that at this level DSE is unstable due to very small population counts. Therefore, the 'DSE counts' for the sampled postcodes within

each cluster of postcodes are constrained to sum to the ‘DSE count’ calculated for the cluster. (The cluster level is chosen to be the constraint and not the index level as this is a compromise between having a small population such that the DSE assumptions are not seriously violated while having large enough counts so that the ‘DSE counts’ are stable.)

4.2 Ratio Estimates

For the second stage of estimation the adjusted ‘DSE count’ (or ratio) for each sampled postcode is then used as the ‘dependent’ variable in a zero-intercept regression model, which links this count with the census count for that postcode. This ratio model is based on the assumption that the 2001 Census count and the dual system adjusted CCS count within each postcode are proportional to each other. Given that it is known from the 1991 Census that undercount varies by age and sex as well as by local characteristics, a separate ratio model within each age-sex group for each HtC category within each Design Group is used. Let Y_{id} denote the adjusted CCS count for a particular age-sex group in postcode i in HtC group d in a particular Design Group, with X_{id} denoting the corresponding 2001 Census count. Estimation in the CCS will be based on the simple ratio model:

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

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

Substituting the least squares estimator for θ_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 Group is:

$$\hat{T} = \sum_{d=1}^3 \left\{ T_{Sd} + \sum_{i \in R_d} (\hat{\theta}_d X_{id}) \right\} = \sum_{d=1}^3 \hat{T}_d \quad (2)$$

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 Group; and R_d is the set of non-sampled postcodes in category d of the HtC index in the Design Group. Strictly speaking the simple model specified by (1) is known to be wrong. The zero covariance assumption in (1) ignores correlation between the cluster of postcodes sampled within a ED. However, the simple least squares estimator (2) remains unbiased under this type of mis-specification, and is only marginally inefficient (Scott and Holt, 1982).

There are two more problems that effect the robustness of the simple model specified by (1). The first problem is the existence of postcodes with a zero count in the census and a non-zero count in the CCS for a particular age-sex group which will induce a positive bias into the ratio estimator. This is dealt with by separately estimating the population total of postcodes with a zero census count for a particular age-sex HtC group using a simple expansion estimator estimated from the CCS postcodes with a zero census count. This is then added to the population total derived for postcodes with a non-zero census count from the ratio estimator. The second problem is when it is necessary to predict for non-sampled postcodes in (2) outside the range of census counts observed in the sample. Again this can lead to a positive bias. This is dealt with by adjusting the ratio model so that θ_d reduces to one when making predictions for such non-sampled postcodes. Further details of these adjustments to the ratio model are given in ONS(ONC(SC))00/03A.

4.3 Variance Estimation

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). In addition the estimator has been adjusted to account for other problems outlined above. Consequently, as the postcodes are clustered within EDs, it is proposed that the conservative ultimate cluster variance estimator will be used. This is given by

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

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 . This estimator is still to be tested using the full estimation methodology outlined above. However, Brown *et al* (1999) presents simulations for a slightly simpler approach and in that case the estimator given by (3) performs well.

4.4 History

The use of dual system estimation evolved early in the project. Initially it was hoped to take advantage of the efficiency gains from using three or more lists. However, the assumption that an individual on any list must be a 'real' person can lead to a large over-enumeration and the lack of suitable individual level lists meant that this possibility was rejected on 27 November 1997.

The use of a combined DSE/regression estimator to make Design Group estimates was proposed and endorsed on 13 November 1998. Subsequent research to address the issue of zero counts has led to the proposal in this paper, which the Steering Committee will be asked to endorse on 9 February 2000.

5. Local Authority District Estimation

Section 4 described the methodology for producing direct estimates by age and sex for each Design Group in the UK. In the case of a LAD with a population of around 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 Groups this will not be the case. For these smaller LADs it will be necessary to carry out a further estimation step, and allocate the Design Group estimate to the LADs constituting this area.

5.1 Small area estimation

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 Group) 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 involves the introduction of small area specific effects 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). An investigation of the different types of approaches that could be used indicated that either a simple synthetic or one which made an adjustment for each LAD to the synthetic estimator should be used. Although the simple synthetic approach has the better precision when the LADs constituting the

Design Group are relatively homogeneous with respect to the structure of their census response rates, this is not the case when large LAD effects are present. Therefore, it is recommended that a LAD adjusted synthetic estimate should be adopted to provide a more robust methodology. This research is contained in ONS(ONC(SC))00/03B.

5.2 Model for estimation

As described in the previous section, direct estimation at the CCS Design Group is based on a simple ratio model 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 Group by including a fixed LAD effect. The LAD adjusted synthetic model used is one that includes an overall age-sex effect (defined at a set of collapsed age-sex groups category level) and an LAD specific effect to distinguish between the LADs. These LAD effects are assumed to cancel out at Design Group level. The approach is implemented separately for each HtC index strata within a Design Group. Let Y_{iadl} denote the adjusted CCS count for a particular age-sex group a in postcode i within HtC strata d of LAD l , with X_{iadl} being the corresponding 2001 Census count. We let c represent the collapsed age-sex groups. The model specification underpinning this approach is:

$$Y_{iadl} = (\theta_{cd} + \gamma_{dl})X_{iadl} + \varepsilon_{iadl}\sqrt{X_{iadl}}; \quad \text{for } a \in c \text{ and } i \in d$$

$$\text{Var}(Y_{iadl} | X_{iadl}) = \sigma_d^2 X_{iadl}$$

$$\text{Cov}(Y_{iadl}, Y_{jbem} | X_{iadl}, X_{jbem}) = 0 \text{ for all } i \neq j$$

with Estimator

$$\hat{T}_{al} = \sum_{d=1}^3 \left\{ T_{Sadl} + \sum_{i \in R_{dl}} (\hat{\theta}_{cd} + \hat{\gamma}_{dl}) X_{iadl} \right\} \text{ for } a \in c.$$

where T_{Sadl} is the adjusted age-sex group a CCS count for the sampled postcodes within HtC category d of LAD l ; and R_{dl} is the set of nonsampled postcodes in category d of the HtC index within LAD l .

The requirement that LAD effects cancel out at the Design Group level is implemented by imposing the constraint $\sum_{l \in G} \gamma_{dl} = 0$. This means that we are fitting an overall Design Group age-sex slope parameter, and then making an adjustment to this slope to take account of the differences between the LADs.

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

5.3 History

This strategy has evolved from papers presented at the Leeds Workshop in May 1998. It will be presented for endorsement by the Steering Committee on 9 February 2000.

6. Demographic Estimates and Quality Assurance

While the 2001 Census based ONC estimates will be considered as the 'Gold Standard' it is important that there is a quality assurance process. This QA process is under development following the strategy laid out in ONS(ONC(SC))00/04 Central to the QA process is the use of the best possible comparable demographic estimates as well as data from other administrative sources which can serve as an independent check on the plausibility of ONC estimates.

6.1 Demographic Estimates

Demographic Estimates will be made for 2001 by 'rolling forward' information from the 1981 Census, using registration data on births and deaths, and migration information from a number of sources. Different levels of error are associated with these sources. Thus in year t the population P_t is given by:

$$P_t = P_0 + \sum_i (B_i - D_i + I_i - E_i),$$

where P_0 is the base population and B, D, I and E are respectively the Births, Deaths, Immigrants and Emigrants in each subsequent year.

There will be a plausibility range around all population estimates and current work is investigating this. Two strategies are being investigated:

- Using advice from an independent panel of experts, upper and lower variants of the national population are being estimated. These will include variants on levels of fertility, mortality and migration as well as on the hard to count groups such as refugees and armed forces;
- At a sub-national level, examining a method that uses the assumption that for similar types of areas, errors between 1991 and 2001 will broadly be constant to those observed between 1981 and 1991.

6.2 Administrative records

The Demographic Estimates make some use of the higher quality Administrative Registers and provide the best plausible single comparators for QA purposes. However administrative records may provide important aggregate level comparators for specific age groups. The availability, reliability and quality of these data sources are currently being investigated within the ONS.

For example the Department of Social Security data on the number of Retirement Pension and Child Benefit claimants. This administrative source is believed to offer almost complete coverage of the elderly and of young children - these two groups have been relatively poorly enumerated in past censuses.

6.3 The QA Process

The QA process will comprise demographic analyses and qualitative judgements. The process will be as follows:

1. Demographic analyses of ONC estimates of population will be undertaken at all three levels of aggregation. These will include analyses of sex ratios and dependency ratios.

2. At specific age groups certain administration records will prove robust. For example birth registrations provide a base from which to form comparators for the very young and at a national level pensions data are likely to provide useful comparators.
3. The distributions of absent households and estimated CCS response/nonresponse rates will be examined.
4. Broad comparisons will be drawn with (a) the demographic estimates from ONS Population Estimates Unit together with the plausibility ranges on which work is progressing (b) estimates of special groups such as Armed Forces Personnel.

6.4 Comparisons with GRO(S)

GRO(S) are proposing a broadly similar strategy. However, because it is the belief of GRO(S) that administrative records in Scotland present a better opportunity for comparison through improved population estimates than in other parts of the UK research is being undertaken to develop a strategy to prepare improved population estimates for comparison with ONC based estimates.

6.5 History

The initial strategy for the ONC presented recognised the need to use demographic estimates and possible use administrative records as a check on the ONC estimates and to identify the best sources for comparison.

- The November 1997 meeting the Committee endorsed the use of the 1981 adjusted Census results as the best rolled forward estimates to benchmark the 2001 adjusted Census results.
- In April 1998 the Steering Committee agreed that cohort analyses should not be pursued at the sub-national level and that a panel of experts to be used to provide plausibility ranges around the demographic estimates.
- Work into the calculation of plausibility ranges for national demographic estimates was presented to the Steering Committee in July 1999. Where it was agreed to there needed to be further work to into disaggregation by age and sex.

The quality assurance strategy will be presented for endorsement by the Steering Committee on 9 February 2000.

7. ONC Imputation & Weighting

7.1 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. The probabilities are then inverted to form coverage weights which are calibrated to agree with the total population estimates by age-sex group and by household size in each LAD. These calibrated coverage weights form the basis of a donor

imputation system which creates synthetic households and individuals to compensate for those estimated to have been missed by the Census.

The modelling of census coverage underlying this procedure is based on the fact that there are two ways in which individuals can 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.

7.2 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 ONC estimation process and the final adjusted database is constrained to satisfy these estimated totals at both the Design Group and the LAD 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 ONC estimates for the LAD. Consequently, these weights are calibrated, using an iterative scaling procedure, to ensure these constraints are satisfied.

7.3 Creating Individual Coverage Weights

Coverage weights for individuals counted by the Census are obtained using similar assumptions to those described above. 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 violated in practice, the extra people are again accounted for by constraining to the ONC estimated totals at the LAD level. Using these assumptions it is only necessary to consider individuals in households counted by both the Census and the CCS. In this case the possible categories are:

- a) Counted by the Census, but missed by the CCS;
- b) Counted by the CCS, but missed by the Census;
- b) Counted by both the Census and the CCS.

Again, matched Census/CCS data and an assumed multinomial logistic model are used to estimate the probability $\pi_{kjil}^{(r)}$ that individual k in household j in postcode i in HtC group d in LAD l has outcome r . As with the household model the individual probabilities for outcomes $r = a$ and $r = c$ depend on individual and household characteristics as measured by 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

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

can be calculated.

7.4 Donor Imputation for Missed Households

This stage of the process uses the household weights to impute households completely missed by the Census. 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 synthetic household is imputed near the location where this event takes place. The donor household for this imputation is defined on the basis of the characteristics of the households with the ‘current’ weight and not only donates the household characteristics but all the individuals within the household. This process ensures that, after the imputation of missed households, the total number of households matches the ONC estimated LAD total. It will also match on totals defined by any other variables to which the household weights have been calibrated.

7.5 Donor Imputation for Missed Individuals

This is the most complex stage of the imputation process since adding individuals to households changes the structure of the recipient household. This stage is best considered in two parts. The first identifies how many individuals need to be imputed and obtains the appropriate donors. 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 ‘current’ characteristics define the basic characteristics of that person. A donor household is then located which contains a person of the required type. Second, the person is imputed into a ‘nearby’ recipient household. The recipient household is the household nearest to the donor household in terms of both space and household structure. The imputed person is added into the recipient household. The recipient household is then subject to Census edit checks to ensure internal consistency.

7.6 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 an addition to a counted household. Typically, this results in an excess of synthetic individuals on the database. 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 resulting household size distribution is correct. A process of ‘pruning off’ and ‘grafting on’ imputed individuals from the database is then carried out until these key LAD totals are achieved.

Eventually, an individual level database will be created which will represent the best estimate of what would have been collected had the 2001 Census not been subject to underenumeration. Tabulations derived from this database will automatically include compensation for underenumeration and therefore all add to the 'One Number'.

Further details on the imputation process are given in ONS(ONC(SC))99/08.

7.7 History

The initial strategy was presented first to the Steering Committee on 12 June 1997. At the Steering Committee on 27 November 1997 it was agreed to investigate both weighting and imputation. Imputation became the favoured option following consultation at the Leeds workshop and was endorsed on 13 November 1998. The imputation strategy has subsequently been refined and was agreed on 1 July 1999. Remaining research is investigating the finer details of pruning and grafting together with work on the imputation of the relationship matrix (this requires Dress Rehearsal data).

8. References

Charlton, J., Chappell, R. and Diamond, I. (1998). Demographic Analyses in Support of a One Number Census. *Proceedings of Statistics Canada Symposium 97*, 51-57.

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.

ONS(ONC(SC))98/12 - Census Coverage Survey: The precision of population estimates for different sample sizes and design areas.

ONS(ONC(SC))98/14 – One Number Census Matching.

ONS(ONC(SC))99/05 – Uncertainty Intervals for National Demographic Estimates.

ONS(ONC(SC))99/08 – A donor imputation system to create a census database fully adjusted for underenumeration.

ONS(ONC(SC))00/03A – Estimation Strategy for Design Group Estimates by Age and Sex from the Census Coverage Survey

ONS(ONC(SC))00/03B – One Number Census Local Authority Estimation.

ONS(ONC(SC))00/04 – A Quality assurance and Contingency Strategy for the One Number Census.

ONS(ONC(SC))00/10 – Design Groups for 2001.

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.