

**ONE NUMBER CENSUS STEERING COMMITTEE****One Number Census Methodology**

This paper describes the ONC methodology for 2001. The paper focuses on the approach for England and Wales. The ONC in Northern Ireland is described in more detail in Annex A, and the ONC in Scotland is described in more detail in Annex B.

**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**

**January 2001**

# The Methodology for Achieving a One Number Census in 2001

## 1. Background

One of the major uses of the decennial census is to provide 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.

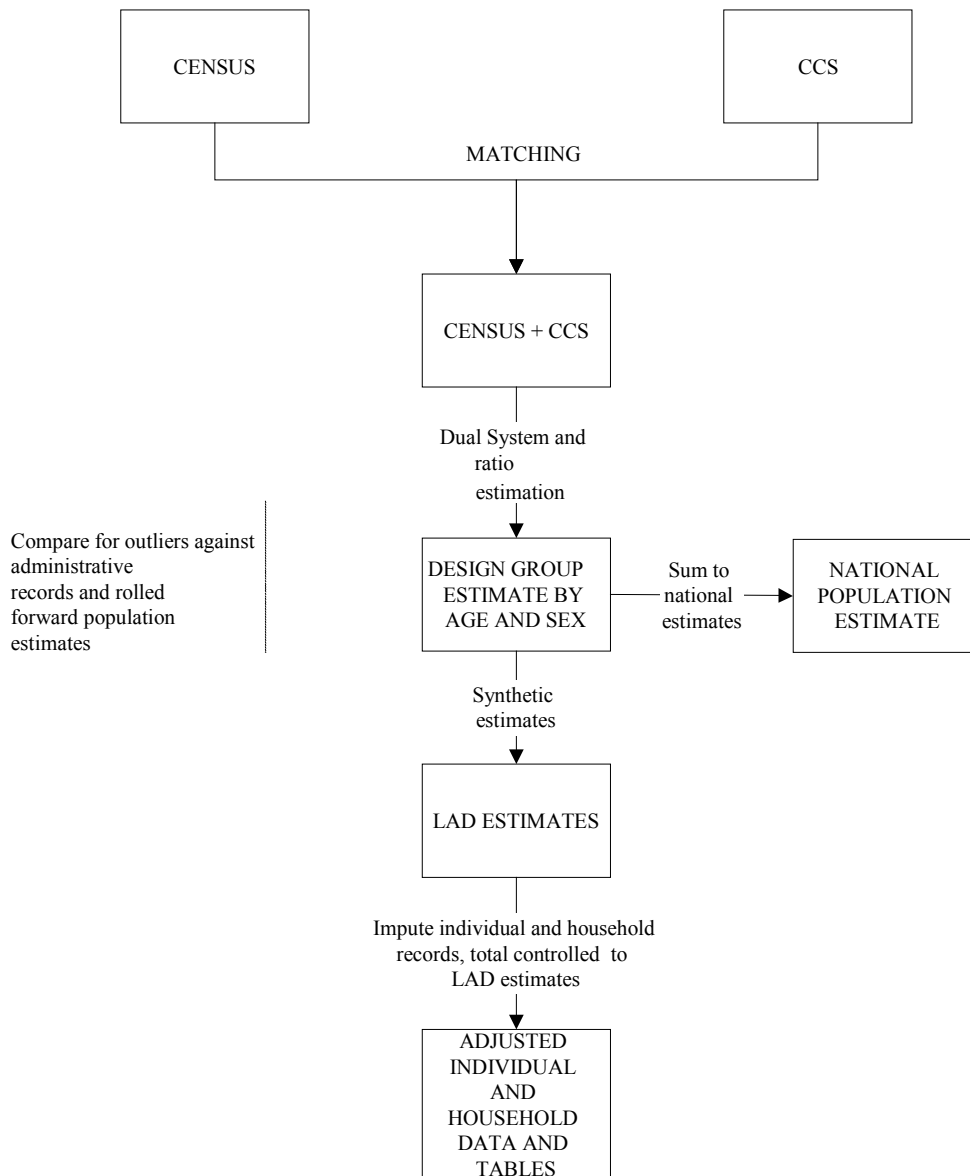
Despite efforts to maximise coverage in the 2001 Census, it is only realistic to expect there will be some degree of underenumeration. The One Number Census (ONC) project aims to measure this underenumeration, provide a clear link between the Census counts and the population estimates, and adjust all 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

- a) A Census Coverage Survey (CCS) will re-enumerate a sample of postcodes. The survey will collect data on a small number of key variables central to measuring underenumeration.
- b) The CCS data will be matched, using a probability based matching procedure, against individual Census records.
- c) 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 England and Wales. 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. There are 376 LADs of varying population sizes.
- d) LAD estimates will be derived from the Design Group estimates using synthetic estimation.

- e) 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.
- f) 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**



## 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 .

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 employ 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, one would expect 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 EDs within each Design Group are stratified by a national 'Hard to Count' (HtC) score. This score is calculated by combining 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 HtC score is an extended version of that used in the 1999 Rehearsal. It is based on the following variables from the 1991 Census:

- proportion of unemployed persons;
- proportion of imputed households;
- proportion of persons whose country of birth is non English speaking;
- proportion of households in multiply-occupied buildings; and
- proportion of households which were privately rented.

The score is derived by a simple summation of the above proportions. 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). The research undertaken to determine the make up and distribution of the index is described in ONS(ONC(SC))00/15 and ONS(ONC(SC))01/06. In particular, consideration was made of a distribution which weighted more of the sample into the hardest to count EDs, for example by choosing a 70%, 20%, 10% distribution. The empirical results indicate that such a distribution does no better than the 40% 40% 20% distribution which is preferred on two grounds. First, there is a necessity to base the score on 1991 Census data - clearly some areas will have changed over time and a more uniform distribution will minimise any biases thus caused. Second, the improvements in Census collection procedures will most likely lead to reductions in underenumeration in some areas but they may be compensated by increases in underenumeration in other areas which have previously been well enumerated - an example may be the increasing number of single person households who can, traditionally be difficult to enumerate.

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 for each age-sex group of interest, within the strata defined by the HtC index and by size ranges corresponding to 1991 Census counts, the true 2001 ED population counts will be independently and identically distributed. The allocation of the sample of EDs between the size strata is then designed to minimise the sampling variability of a stratified expansion estimate of the Design Group strata 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 and distribution**

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 (a number of censuses and associated CCSs were simulated using 1991 census data), 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 and subsequent work in ONS(ONC(SC))01/03.

It is expected that the underenumeration in 2001 will not be evenly spread across the country. Therefore, it is sensible to weight the sample towards the areas that are expected to have a high undercount. The aim is to produce Design Group estimates with comparable accuracy. Therefore the amount of the sample allocated to each strata must be that which gives a similar expected precision. The actual obtained precision will be dependent on the population size of the Design Group, the level of underenumeration in the 2001 Census and the CCS sample size. Within each Design Group it is expected that the variance will be higher in the hardest to count EDs and hence there will be a larger sample size in the hardest to count areas.

The final allocation of the sample resulted in 4.5% of HtC category 3 EDs (the hardest) included in the sample compared to 3.4% of HtC category 1 EDs (the easiest). Therefore, a Design Group made up of mostly hard to count EDs will be allocated a larger sample size than a Design Group made up of easy to count areas. For example, the Design Group containing Brent and Haringey (population 474,000) has a sample size of 41 EDs. The Design Group containing Caerphilly, Merthyr and Taff UAs (population 466,000) has a sample size of 33 EDs. This reflects the relative distributions of hard to count scores within these two Design Groups.

It must be noted that there must be a balance between weighting towards the harder to count areas and producing a robust strategy. One can only guess the distribution of the undercount, and therefore one must have a design that will provide estimates of an acceptable precision if predictions are inaccurate.

## **2.5 History**

At each stage of the paper a short historical section will provide a description of the processes through which the development of that stage went. The CCS 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 it permitted an independent assessment of all aspects of the enumeration.
- c) The efficiency of the design and decisions regarding the sample size were approved on 13 November 1998.
- d) The makeup of the Design Groups was agreed by the ONC Project Board on 20 December 1999.
- e) The composition of the Hard to Count Index was agreed at the Steering Committee meeting 28 June 2000.

### **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 (for example postcode) for an initial grouping of the data to reduce the number of comparisons.
2. Automatically match households within the groups defined by the blocking variables using exact and probability matching.
3. Automatically match individuals within matched household pairs using exact and probability matching.
4. Review a sample of matched households and individuals to confirm accuracy.
5. Present low probability pairs to matchers for clerical resolution.

6. Re-block unmatched data (for example by EA and phonetic coding of surname of head of household).
7. Present pairs, ranked by matching weights, to matchers for clerical resolution.
8. Expert matchers review a sample of clerically matched records.
9. Clerical search for unmatched CCS records with all blocking removed. Search at both household and individual level. Refer marginal cases to expert matchers.
10. Expert matchers review a sample of stage 9 clerical matches for quality assurance.
11. Second, independent search for unmatched CCS records remaining. Search performed by expert matchers. Search at both household and individual level.

More details of the proposed matching methodology are given in ONS(ONC(SC))01/05 and 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. An initial evaluation was presented to the Steering Committee on 28 June, since then further evaluation has been carried out and is reported in ONS(ONC(SC))01/05.

## **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 (the cluster is defined as the postcodes selected within each PSU) are constrained to sum to the DSE count calculated for the cluster. The cluster level is chosen to be the constraint and not the postcode 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:

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

$$\text{Cov}\{Y_{id}, Y_{jf} | X_{id}, X_{jf}\} = 0 \text{ for all } i \neq j$$
(1)

Substituting the least squares estimator for  $\theta_d$  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:

$$\hat{T} = \sum_{d=1}^3 \left\{ T_{Sd} + \sum_{i \in R_d} (\hat{\theta}_d X_{id}) \right\} = \sum_{d=1}^3 \hat{T}_d$$
(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  $\theta_d$  is reduced 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 and ONS(ONC(SC))00/16.

### 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, as outlined above, to account for other problems. Consequently, as the postcodes are clustered within EDs, it is proposed that the ‘drop one PSU’ jackknife variance estimator will be used. This is given by:

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

where  $\hat{T}_d^{(e)}$  denotes the BLUP for the population total of category  $d$  of the HtC index based on the sample data excluding data from ED  $e$ . Earlier work on variance estimation in Brown *et al* (1999) used the ultimate cluster variance estimator with a simpler estimation strategy. However, the simulations in ONS(ONC(SC))00/16 have shown that with the more complex estimation strategy outlined above, the estimator given by (3) performs well while the ultimate cluster variance estimator is not as good.

### 4.4 History

The use of dual system estimation developed 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 where there are people on lists who should not really be on the list. A 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 issues of zero counts and prediction outside of the range has led to the proposals in this paper, which the Steering Committee endorsed at meetings on 9 February 2000 and 28 June 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 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, to estimate the population of the LADs constituting the particular Design Group.

## 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 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 estimate 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, a LAD adjusted synthetic estimate will 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 level 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 categories 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:

$$\begin{aligned}
 Y_{iadl} &= (\theta_{cd} + \gamma_{dl})X_{iadl} + \varepsilon_{iadl}\sqrt{X_{iadl}}; \quad \text{for } a \in c \text{ and } i \in d \\
 \text{Var}(Y_{iadl} | X_{iadl}) &= \sigma_d^2 X_{iadl} \\
 \text{Cov}(Y_{iadl}, Y_{jbem} | X_{iadl}, X_{jbem}) &= 0 \text{ for all } i \neq j
 \end{aligned} \tag{4}$$

with estimator

$$\hat{T}_{al} = \sum_{d=1}^3 \left\{ T_{Sadl} + \sum_{i \in R_{dl}} (\hat{\theta}_{cd} + \hat{\gamma}_{dl}) X_{iadl} \right\} \quad \text{for } a \in c. \tag{5}$$

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 one is 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  $\gamma_{dl}$  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 was endorsed 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 (QA) process. The strategy is laid out in ONS(ONC(SC))00/04 and Census Advisory Group paper (00)16. 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:

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

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 diagnostic range around all population estimates. Two strategies will be used:

- At a national level, using advice from an independent panel of experts, upper and lower variants of the national population have been estimated. These include variants on levels of fertility, mortality and migration as well as on the hard to count groups such as refugees and armed forces (see ONS(ONC(SC))99/05);
- At a sub-national level, the ONC estimates will be compared with a portfolio of estimates from both demographic and administrative sources.

## **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 will provide important aggregate level comparators for specific age groups.

An example is 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 age 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. At a sub-national level use will be made of child benefit and pensions data if their accuracy can be demonstrated.
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 diagnostic ranges on which work is progressing (b) estimates of special groups such as Armed Forces Personnel.

## **6.4 History**

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

At 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 be used to provide diagnostic ranges around the demographic estimates.

Work into the calculation of diagnostic ranges for national demographic estimates was presented to the Steering Committee in July 1999, where it was agreed to carry out further work into disaggregation by age and sex.

The quality assurance strategy was endorsed by the Steering Committee on 9 February 2000. Further refinements of the strategy and proposals for taking it forward were presented to the Steering Committee on 28 June 2000 (see ONS(ONC(SC))00/18).

Members of the six Census Advisory Groups and the Liaison Group on Population Statistics (LGPS) have been consulted on the latest plans for quality assurance of ONC estimates (Census Advisory Group paper (00)16).

## **7. ONC Imputation**

### **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  $\pi_{kjidl}^{(r)}$  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

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

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. The final details of pruning and grafting together with the use of Dummy forms was presented to the Steering Committee on 28 June 2000.

## 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.

Office for National Statistics Census Advisory Group paper (00)16 - A Quality Assurance & Contingency strategy for the One Number Census. Titchfield: ONC.

*Office for National Statistics One Number Census Steering Committee Papers*

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

ONS(ONC(SC))98/14 - One Number Census Matching. Titchfield: ONC

ONS(ONC(SC))99/05 - Uncertainty Intervals for National Demographic Estimates. Titchfield: ONC

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

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

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

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

ONS(ONC(SC))00/10 - Design Groups for 2001. Titchfield: ONC

ONS(ONC(SC))00/15 - 2001 Hard to Count Index. Titchfield: ONC

ONS(ONC(SC))00/16 - Estimation Update. Titchfield: ONC

ONS(ONC(SC))01/03 – Census Coverage Survey sample size, Coverage and the impact of dependence on the One Number Census. Titchfield: ONC

ONS(ONC(SC))01/05 – ONC Matching. Titchfield: ONC

ONS(ONC(SC))01/06 – Transformation of the Hard to Count variables. Titchfield: ONC

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.

## **The One Number Census in Northern Ireland**

### **1. Introduction**

In recent decades the Census of Population in Northern Ireland has followed broadly the methodology and questionnaire content of the other Censuses conducted in the rest of the UK. However, separate output has traditionally been produced for the Northern Ireland Census with no outputs for the UK as a whole. At the start of the planning process for the 2001 Census of Population, NISRA decided to become more fully integrated with the Censuses proposed for England and Wales and Scotland, with the ultimate aim of producing UK-level output. Accordingly, NISRA plan to apply the One Number Census principles, adjusting 2001 Census outputs for estimated under-enumeration, and producing Census output which is conceptually consistent with the mid-year population estimates.

The One Number Census methodology in Northern Ireland will broadly follow that in Great Britain, as described in many papers produced by the ONC Project Board and Steering Committee. This paper will take the previous methodology papers as given and concentrate on any differences that will apply in Northern Ireland; it is believed that these differences are marginal.

### **2. Is there a need for a One Number Census approach in Northern Ireland?**

It is generally accepted that a successful census was conducted in Northern Ireland in 1991. However, even with a 'successful' census, the follow-up Census Validation Survey to the 1991 Census identified under-enumeration of 0.66 per cent although the survey was too small to allow separate estimates of under-enumeration by age (1991 Northern Ireland Census General Report, NISRA). Accordingly, at that time the mid-year estimates (MYEs) of population took the 1991 Census base and grossed-up by 0.66 per cent (over and above correction factors for HM Forces, moving students to term-time address and changes between Census Day and 30 June).

Later analyses which retrospectively examined administrative data sources suggested that there were a number of problems, not identified by the Census Validation Survey, for, in particular, very young children and the very elderly (NISRA Occasional Paper Number 12, 1999). It is now believed that total under-enumeration in 1991 was of the order of just over 1 per cent but concentrated in certain age-groups.

NISRA's view of the One Number Census is to accept that some under-enumeration is inevitable and to plan accordingly. The aim remains good coverage in the Census, and the One Number Census is viewed as contingency for not achieving full coverage.

### **3. Differences in the 2001 Census in Northern Ireland compared to Great Britain**

The Census methodology in Northern Ireland will be very similar to that in the rest of the UK and NISRA participates in a range of formal Working Groups and Steering Groups, in addition to ad hoc liaison with ONS and GROS colleagues, to ensure consistent practice throughout the UK. Differences in fieldwork practices are minor; some examples are that Northern Ireland will be using an advance letter, and the split of duties between enumerators and Team Leaders may vary between countries. Details are still being finalised, but NI may conduct more stringent completeness checks than England and Wales. These differences are considered to be marginal.

The questionnaire in Northern Ireland is very similar to that in GB. The main differences are:  
Northern Ireland has a religion question (as it has always had);  
The ethnic question in Northern Ireland is less detailed reflecting local need;  
Northern Ireland has an Irish language question, similar to questions in Scotland and Wales;  
An additional household question on the number of floors occupied.

### **3.1 The Census Coverage Survey**

As in GB, ONC methodology will be based on a large-scale face-to-face household survey – the Census Coverage Survey (CCS) - that will commence approximately three weeks after Census Day. CCS fieldwork dates are similar in Northern Ireland and GB.

#### **3.1.1 CCS - questionnaire**

For ONC purposes, the religion question is viewed as more important in Northern Ireland than the ethnic origin question as numbers of people from ethnic minorities are small in Northern Ireland. Accordingly, the CCS questionnaire in Northern Ireland will ask about religion and not ask about ethnic origin. Apart from this the Northern Ireland questionnaire is similar to the GB questionnaires.

#### **3.1.2 CCS - fieldwork**

One of the key differences between the Northern Ireland CCS and its GB counterparts is the use of an experienced fieldforce in Northern Ireland. NISRA's equivalent of Social Survey Division (SSD) in ONS, the Central Survey Unit (CSU), maintains a fieldforce of over 100 interviewers to conduct the Northern Ireland elements or equivalents of the Labour Force Survey, the General Household Survey, the Family Expenditure Survey and a range of other regular and ad-hoc sample surveys. NISRA used the CSU fieldforce to conduct the 1997 Census Test, the follow-up Census Test Evaluation Survey and the 1999 Census Rehearsal. NISRA have decided to use the CSU fieldforce to conduct the CCS.

Fieldwork procedures will generally be similar to those in GB, and joint training is being organised with colleagues in GROS. The biggest difference in the fieldwork procedure (compared to England and Wales) is likely to be that NISRA CCS interviewers will not share workloads in pairs.

#### **3.1.3 CCS - sample size and design**

A fuller description of the proposed Northern Ireland design is given in the technical annex, the content of which is summarised below.

In GB, the CCS is designed as a series of Design or Estimation Areas (EA), where each EA has a population of about 500,000 and each EA has a separate CCS. This 0.5m building block principle has been applied in Northern Ireland (total population 1.7m) resulting in 3 EAs. For administrative purposes, Northern Ireland is split into 26 Local Government Districts (LGDs) - the equivalents of LAs in GB although Northern Ireland's LGDs are generally much smaller (typical population of about 50,000) than LAs in GB. Each EA in Northern Ireland is the aggregate of a number of LGDs, similar to the aggregation of LAs in GB.

Belfast is the largest LGD in Northern Ireland (1998 MYE 287,500) but surrounded by dormitory towns, which are socio-economically different to Belfast itself. Outside of greater Belfast, most of Northern Ireland is very rural by British standards. Northern Ireland's three EAs are thus Belfast (single LGD,

population 287,500), East of Northern Ireland (12 LGDs surrounding Belfast, population 764,400) and West of Northern Ireland (the remaining 13 LGDs, population 636,600).

For each EA, the CCS is being designed in a similar manner to GB, with 1991 EDs being stratified and then used as primary sampling units from which unit postcodes are selected. The CCS will consist of complete re-enumeration of the sampled postcodes. The main difference in the Northern Ireland approach is the method by which EDs are stratified. The stratification of 1991 EDs in GB is based on a Hard to Count index derived from research into under-enumeration in 1991. Some key GB indicators such as multi-occupancy dwellings and ethnic minorities have very low incidence levels in Northern Ireland and there is no research to support similar effects in Northern Ireland. The stratification of 1991 EDs in Northern Ireland will be based on observed response rates to Northern Ireland's (voluntary) 1997 Census Test, which was designed as a fractional replicate of a 2x2x3 experiment where EDs were the sampling units, classified by predominant religious background (3 levels), urban/rural and deprived/non-deprived. The 12 design strata from the 1997 Census Test have been collapsed to 3 levels of stratification for the CCS, merging strata with similar response rates.

A design has been produced which has the same overall sampling fraction as England and Wales (approximately 3.6 per cent of 1991 EDs as primary sampling units); in Northern Ireland this corresponds to 134 EDs. Within each ED, a sample of postcodes, consisting of on average 5 postcodes and 75 households, will be selected for re-enumeration, giving a total sample of 670 postcodes involving 10,000 households. At this stage, the ONS approach (5 random postcodes) and the GROS approach (continue to sample postcodes until a workload is achieved) are both being considered. The experience of the 1999 rehearsal, which demonstrated a very skewed distribution of postcode size (by number of households), makes the GROS approach attractive.

### **3.1.4 Efficiency of the proposed CCS design**

The efficiency of the proposed design has been examined through a simulation exercise based on the 1991 census database for Northern Ireland. Under the assumptions of census coverage of 97.25 per cent, CCS coverage of 90 per cent (households) and 98 per cent (people within households) and independence for the dual estimator, the simulations suggest a relative RMSE of 0.46 per cent for the Northern Ireland population estimate. Further details including analysis by EA are given in the Technical Annex.

## **4. ONC analysis**

The ONC statistical analysis (matching, use of dual estimators, regression analysis and so forth) is planned to be similar to that in GB except that religion will take the place of ethnic origin. The analysis software is being developed by ONS and discussions are underway about how the ONC analysis will actually be delivered. It is likely that the ONC analyses for all UK censuses will be run on ONS computers; a secure communications link is being set up so that NISRA can have ready access to the diagnostic outputs and be proactively involved in the analysis of Northern Ireland data.

Associated with being relatively small in population terms, Northern Ireland LGDs have very limited powers and NISRA do not anticipate the extent of interest expected from LAs in GB regarding Census estimates of their population. Accordingly, NISRA do not anticipate demand for separate under-enumeration estimation for each LGD. Each EA is however composed of an aggregation of NUTS level 3 units (the NUTS classification is the standard aggregation method used when making statistical returns to the European Union); each of the East and West EAs are the aggregate of 2 NUTS areas. NISRA intend to make estimates for the NUTS areas within each EA in a similar manner to ONS's plan for LAs within EA, that is allowing different patterns of under-enumeration for different NUTS areas within an EA.

Estimates for LGDs will then be produced using standard synthetic estimation for each LGD within the NUTS unit.

NISRA will require age x sex population estimates for each LGD by summer 2002 for the purpose of producing the 2001 mid-year population estimates. These LGD figures must of course be consistent with those that will be produced in census outputs in 2003. Therefore, the 'pruning and grafting' step will be required at LGD level.

#### 5. ONC contingency

Northern Ireland anticipates using similar procedures to those being developed in the rest of the UK, and work is ongoing to develop tolerance estimates by which adjusted Census counts can be compared with existing estimates such as the Mid-Year Estimates of population and administrative data sources.

#### **6. Summary**

In summary, the One Number Census process will be applied to Northern Ireland 2001 census data in a similar way to the rest of the UK to ensure that consistent UK census output can be produced.

## ANNEX A1

### Implementing the CCS design in Northern Ireland

#### 1. Design Groups

The Northern Ireland census count in 1991 was about 1.6 million in 3,729 EDs and the population is projected to be about 1.7 million by 2001, suggesting the use of 3 Estimation Areas for 2001 CCS purposes. The country is split up into 26 Local Government Districts (LGDs) that are similar to Local Authority Districts (LADs) in England and Wales. However, they differ in that they are generally smaller in population terms than those in England and Wales (some are very small, Moyle has a population of 15,000) and in general they do not have as much political power as LADs in England and Wales.

Belfast is easily the largest LGD in population terms, with a population of about 300,000. The LGD boundary defines the urban area very tightly, and many areas that might be considered as suburbs of the city are in neighbouring LGDs such as Castlereagh and Newtownabbey. A consequence of this is that the population of the LGD area is declining as people move to the suburbs. However, it is still important and likely to differ in terms of underenumeration from the rest of Northern Ireland. Therefore Belfast is considered as a design group on its own. The rest of the LGDs are grouped into a standard East and West classification as shown in Table 1.

Table 1: Classification of LGDs (excluding Belfast) into Two Design Groups (Number of EDs is based on the 1991 Census)

East		West	
LGD	Number of EDs	LGD	Number of EDs
Antrim	88	Armagh	153
Ards	162	Ballymoney	71
Ballymena	132	Coleraine	120
Banbridge	92	Cookstown	89
Carrickfergus	64	Derry	182
Castlereagh	118	Dungannon	133
Craigavon	155	Fermanagh	214
Down	146	Limavady	64
Larne	76	Magherafelt	90
Lisburn	202	Moyle	47
Newtownabbey	147	Newry and Mourne	229
North Down	145	Omagh	137
		Strabane	106
Total	1527	Total	1635

This classification means that using 1991 Census figures the population of the Belfast design group is 279,215, the population of the East design group is 700,364, and the population of the West design group is 598,111. This variety in the population size will have consequences when choosing sample sizes.

## 2. Enumeration District Type (EDT) Index for Northern Ireland

In England and Wales, EDs have been classified using a Hard to Count Index, which incorporates local indicators, such as multi-occupancy housing and ethnic minority populations, which research has demonstrated to be related to underenumeration. For Northern Ireland no similar research is available, and at least some of the indicators have very low incidence levels.

The stratification of 1991 EDs in Northern Ireland will be based on observed response rates to Northern Ireland's (voluntary) 1997 Census Test, which was designed as a fractional replicate of a 2x2x3 experiment where EDs were the sampling units, classified by predominant religious background (3 levels), urban/rural and deprived/non-deprived. Further details on the classification methods used in the 1997 Census Test and the observed response rates can be found in NISRA Occasional Paper Number 13 (1999).

The religious background classification was reduced from 3 levels to 2 on the basis of similar response rates, giving eight initial strata for the EDs, as shown in table 2 where the strata are ranked according to observed response rates (highest first).

Table 2: Ranking of the EDT Index

Level of Index	Religion	Location	Deprivation Status
1	Protestant	Rural	Not Deprived
2	Protestant	Rural	Deprived
3	Catholic & Mixed	Rural	Not Deprived
4	Protestant	Urban	Not Deprived
5	Protestant	Urban	Deprived
6	Catholic & Mixed	Urban	Not Deprived
7	Catholic & Mixed	Rural	Deprived
8	Catholic & Mixed	Urban	Deprived

While it is desirable to spread the sample over all the eight categories it will not be possible to estimate independently in all eight. Therefore, estimation will use a three level categorisation that combines the categories in Table 2. Levels 1 to 5 will form an 'easy to count' group containing about 33 per cent of the population, levels 6 and 7 will form a middle group containing about 50 per cent of the population, with level 8 forming a 'hard to count' group. The distribution of the EDT index across the design groups is given in Table 3 for both the full and collapsed categorisation.

Table 3: Distribution of the EDT Index by Design Group

Level of Index	Number of EDs			
	Belfast	East	West	Northern Ireland
1	-	135	61	196
2	-	59	100	159
3	-	149	145	294
4	80	251	38	369
5	128	63	18	209
Estimation 1	208	657	362	1227
6	180	524	275	979
7	-	166	733	899
Estimation 2	180	690	1008	1878
8	179	180	265	624
Estimation 3	179	180	265	624
All	567	1527	1635	3729

### 3. Allocating the Sample at Stage One

The approach that is being used in England and Wales forms the basis of the allocation, with some specific differences. The number of Northern Ireland EDs sampled is specified using the same sampling fraction as for England and Wales (approximately 3.6 per cent). For Northern Ireland this implies a first-stage sample of 134 EDs. An initial allocation to the design group by collapsed EDT index is made proportional to the population sizes within the groups. The use of population at this stage rather than number of EDs reflects the fact that the East design group is the largest in terms of population but has less EDs than the West design group. If any allocation is less than eight EDs this is forced to equal eight and the proportional allocation is repeated for the remaining groups. This is to guarantee sufficient sample within each collapsed EDT index category for estimation. The specified sample of EDs is then proportionally allocated (by number of EDs) to the full EDT index. This will ensure that although the sample is designed for estimation using the specified collapsed categories the sample will be spread across all eight categories and allow estimation within a different set of collapsed categories.

The ONC design assumes a second level of stratification below EDT index based on population size. The problem is choosing which age-sex ED counts to use as the size variable. In England and Wales this has been solved by constructing a design variable based on the first three principal components derived from six age-sex groups (males 0-4, females 0-4, males 20-24, males 25-29, males 30-34, females 85+) that suffered high underenumeration in 1991. The within index stratum boundaries are then defined using minimum variance cluster analysis on the three principal components. Optimal allocation based on the design variable is used to allocate a pre-specified within index sample to the size strata such that the relative standard error (RSE) for the estimate of the design variable total is minimised. (RSE is the standard error of the estimated total expressed as a percentage of the total and is also called the 'coefficient of variation'.)

The same approach has been taken in Northern Ireland. However, proportional allocation, rather than optimal allocation, is used to allocate the specified sample to the within EDT index size strata such that

the relative standard error (RSE) for the estimate of the design variable total is minimised across the whole design. This approach is not necessarily as ‘efficient’ as the ‘optimal’ allocation approach but it does have one advantage, it spreads the sample evenly across all the strata. This is important for two reasons; firstly it is robust when you have little information about the expected pattern of underenumeration, secondly it looks fair to the user if they perceive there to be little information about the expected pattern of underenumeration.

Table 4 presents an indicative design for Northern Ireland based on a simulation dataset constructed from the 1991 Census data for Northern Ireland. It demonstrates the structure of the final design for the CCS. The total first-stage sample is 130 EDs plus four EDs included as outliers based on the size of the six population groups. (This is consistent with the approach used by ONS.) The lower sample in the West reflects the less dense population in that design group compared to the East but in general the proportional allocation has led to a sample that is spread evenly over the province.

Table 4: Indicative First-Stage Design for Northern Ireland

Group	EDT Index	Total Number of EDs	Number of Size Strata	ED Sample Size
Belfast	4	80	2	4
	5	128	2	5
	6	180	2	9
	8	179	4	7
	All levels	567	10	25
East	1	134	2	4
	2	59	2	2
	3	148	1	5
	4	251	3	7
	5	63	1	2
	6	523	7	21
	7	166	3	6
	8	180	3	8
	All levels	1524	22	55
West	1	61	1	2
	2	100	2	2
	3	145	2	4
	4	38	1	1
	5	18	1	1
	6	274	4	7
	7	733	6	20
	8	265	5	13
	All levels	1634	22	50
Northern Ireland	All levels	3725	54	130

#### 4. Allocating the Sample at Stage Two

In England and Wales, simulation work has suggested that a random selection of five postcodes from each sampled ED (or less in situations where the ED does not contain five postcodes) is a good compromise between clustering for cost efficiency and spreading the sample of postcodes for statistical efficiency. The expected household sample will be approximately 75 households but this is subject to considerable variation. The variation then needs to be handled by the field management systems of the CCS.

An alternative approach being investigated by GROS draws a sample of postcodes from within the sampled ED until a target number of households is achieved based on the PAF data on number of addresses. With the target household approach care must be taken to:

- a) guarantee that those carrying out the CCS fieldwork are not aware of expected numbers of households
- b) account for the fact that the PAF is number of addresses and in areas where multiple-occupancy is expected the number of households will be higher
- c) decide what to do when a sampled ED does not contain the target number of households
- d) handle large postcodes, some of which contain over the specified target, and the situation where a randomly selected postcode is 'too big' based on the previous draws.

If these problems can be satisfactorily taken care of, the approach has some attractions for management of workloads in the CCS. Northern Ireland's experience in the 1999 rehearsal was that the distribution of postcodes (in terms of population size) was very skewed with many postcodes having very small numbers of households. The two approaches to second stage sampling are currently being considered for Northern Ireland, but the end-point will be ED samples containing, on average, 5 postcodes and 75 households.

#### 5. Preliminary Simulation Results

The 1991 census database has been used to simulate the outcomes of using the ONC process to estimate the true population following a census with less than 100 per cent coverage. The outcomes from using the CCS sample size and design described above have been simulated using the 1991 census data, with the main results summarised in table 5.

Table 5: Results of simulation exercise

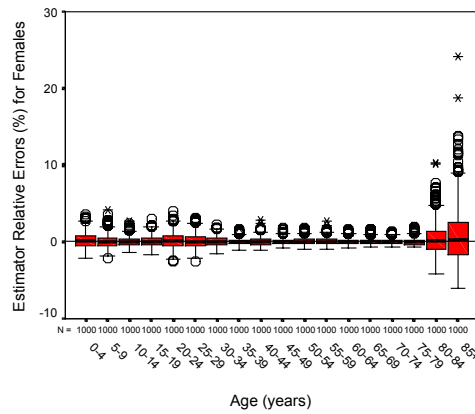
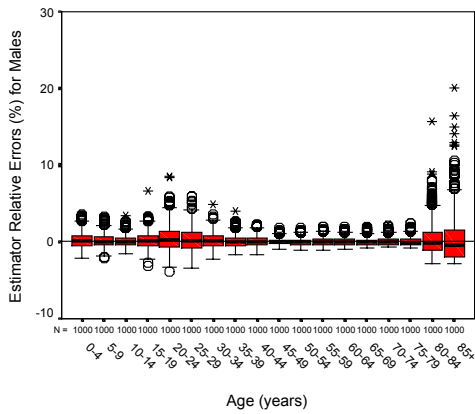
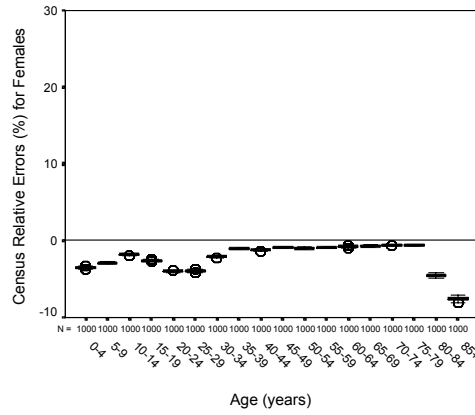
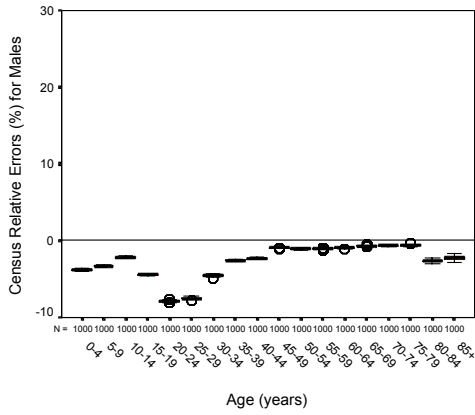
Properties of Estimated Population Total				
Design Group	Census Coverage (%)	Relative Bias (%)	Relative RMSE (%)	Z-Value for Bias
Belfast	92.51	0.19	1.74	3.55
East	98.68	-0.02	0.39	-1.63
West	97.81	0.24	0.75	10.71
Northern Ireland	97.26	0.12	0.46	8.37

The outcome of the simulation is of course largely driven by the underlying assumptions. For the above simulation, the assumptions were

- overall census coverage of 97.25 per cent
- census coverage assumed worst in Belfast
- CCS coverage of 90 per cent of households
- CCS coverage of 98 per cent of individuals within covered households
- Census and CCS independence for dual estimator purposes.

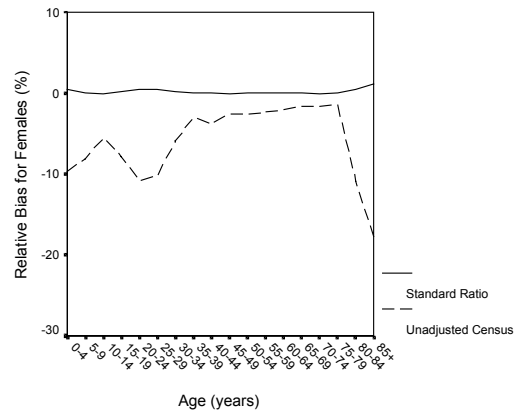
At the Northern Ireland level, the simulation is suggesting a relative RMSE of 0.46 implying a 95 per cent confidence interval of less than 1 per cent. For Belfast, the low census coverage has been recovered (final bias of 0.19 per cent) although with quite a high relative RMSE.

The charts below summarise further results from the simulation. The first charts show how the simulated census underenumeration was distributed by age-group, with high under-enumeration concentrated in people in their 20s and the very elderly. The following charts demonstrate how the ONC process has recovered this underenumeration. The following charts then show the original (census) and final (adjusted) relative bias and relative RMSE for each Design Group (EA) separately.

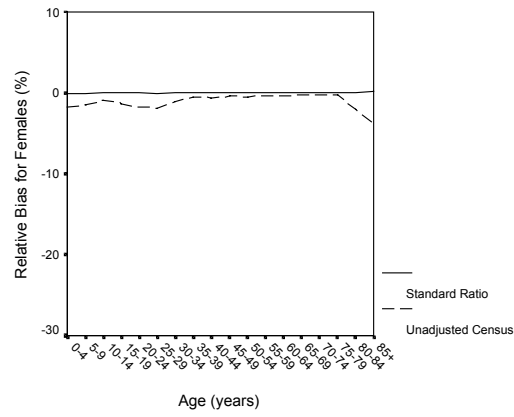
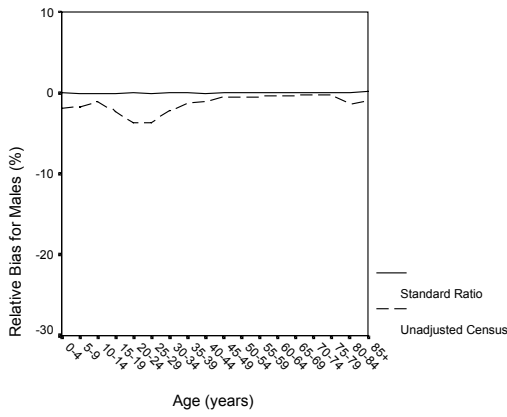


# Relative Bias by Design Group

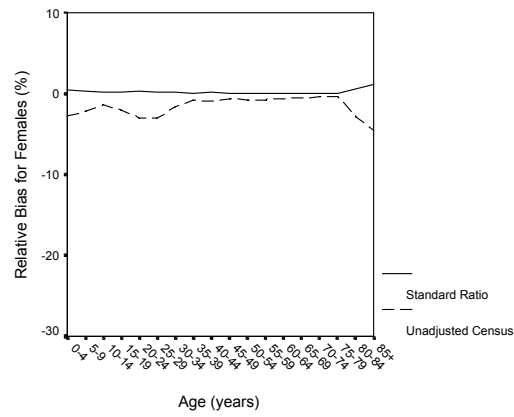
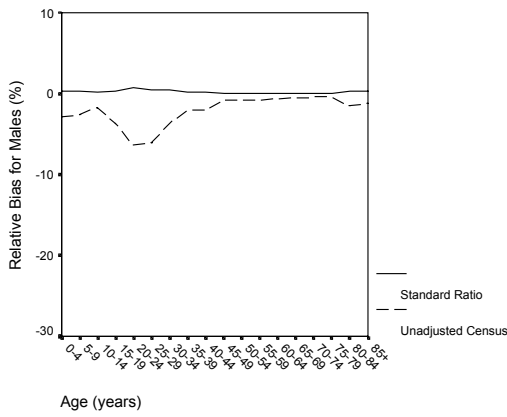
## Belfast



## East

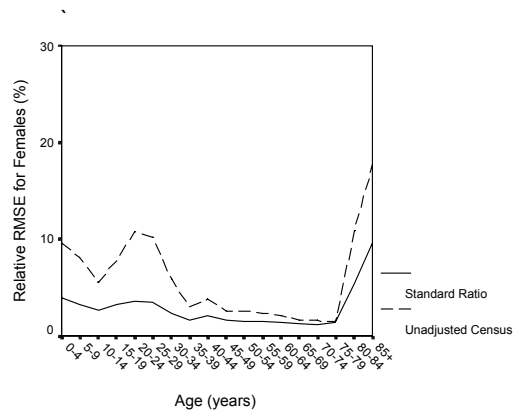
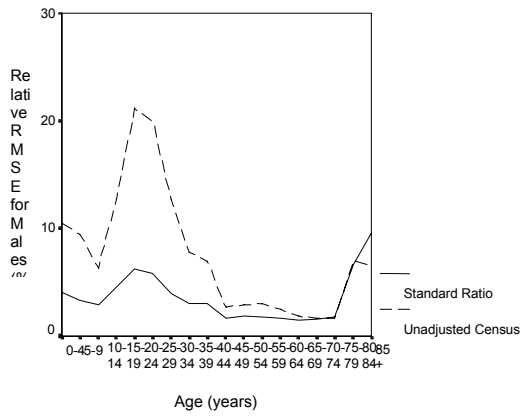


## West

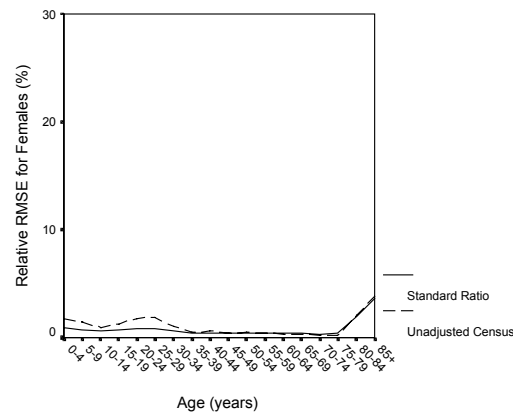
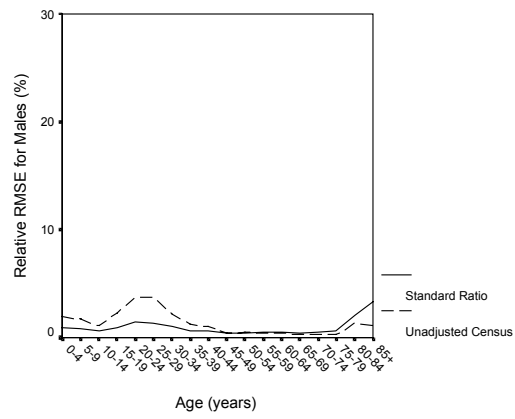


# Relative RMSE by Design Group

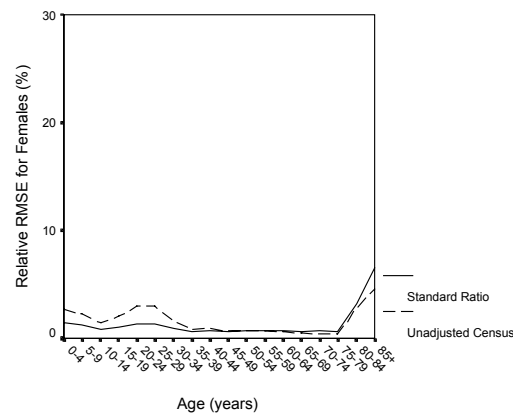
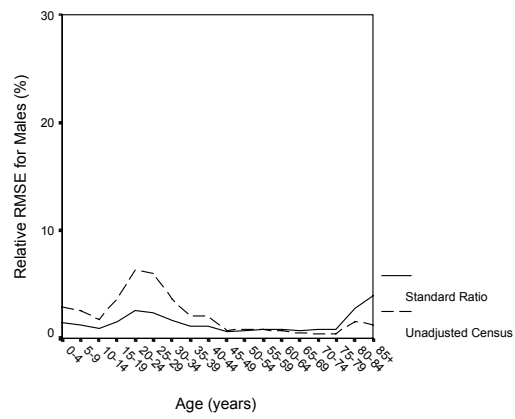
## Belfast



## East



## West



## **ANNEX B**

### **The Methodology for Achieving a One Number Census for Scotland**

**Ian Máté**

#### **1. Background**

The One Number Census methodology in Scotland broadly follows that in other parts of the UK as described in many papers written for the ONC Steering Committee. This paper takes the previous methodology papers as read, especially the ONS paper of the same title as above to which this paper is an annex. It concentrates on the differences in Scotland compared to England and Wales and Northern Ireland. It is believed that these differences are marginal and mainly relate to the design and execution of the Census Coverage Survey.

Most of the ONC development work has been carried out by Ian Diamond and colleagues at Southampton University and Marie Cruddas and Colleagues at ONS; this paper borrows strongly from that work. We would also like to acknowledge the advice received from St Andrew's University staff and Jonathan Ashbridge at GROS.

The Scottish CCS has been run as a separate project from England and Wales and Northern Ireland due to the distinct Scottish Census geography data availability and differences in the physical and social geography of Scotland compared to England.

The methodology of the CCS and the Census has been developed generally for the countries of the UK as a whole in an UK wide project. A full stand-alone paper describing the Scottish CCS and ONC is available.

1.1 GROS shares the view, based on the analysis of the Census results from 1991, that some biased under enumeration in the Census is inevitable. Our aim is to have a Census database adjusted for under enumeration and the One Number Census is, as in other parts of the UK, our contingency for not achieving full cover and for biased under enumeration.

1.2 The One Number Census process, as in ONS, comprises six stages. These include:

- a) A Census Coverage Survey (CCS) will re-enumerate a sample of postcodes. The survey will collect data on a small number of key variables central to measuring under enumeration.
- b) The CCS data will be matched, using a probability based matching procedure, against individual Census records.
- c) 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, in Scotland, are based on Health Board Areas (HBA) and their constituent Council Areas. Council Areas are sometimes split between Health Board Areas. Where they are split, the parts of the Council in each area are treated as if they were Council Areas in their own right. 4 Council Areas are split by Health Board boundaries. Therefore there are 32 Council Areas but 36 'Baileries' (28 Council Areas and 4

x 2 split Council Areas). The 36 Bailerries are referred to as Council Areas below for ease of comparison with the England and Wales methodology.

- d) Council Area estimates will be derived from the Design Group estimates using synthetic estimation.
- e) National, Design Group, Health Board Area and Council Area estimates will be compared with a set of 1991 based estimates to assess their plausibility. In the event that an estimate is implausible, a contingency strategy will be used.
- f) Individual and household level records will be imputed for those estimated to have been missed by the Census.

## **2. The Design of the Census Coverage Survey**

2.01 The aim of the CCS following the 2001 Census is to facilitate the estimation of under enumeration by age and by sex for all Council Areas (CAs) in Scotland. However, a CCS Design with the objective of producing direct estimates for CAs would lead to a prohibitively large sample size. Therefore, to allow a more efficient sampling strategy, geographically contiguous HBAs were aggregated to form Design Groups of about 500,000 people. The optimum population size of Design Groups was investigated in ONS(ONC(SC))98/12 and the actual groupings for Scotland are shown in Table 1.

2.02 These Design Groups are then used independently throughout the whole ONC process as strata for design, estimation and imputation. Within the imputation, there is, however, a slight difference where a donor is sought to fill the details of an imputed person, household or person within household. Processing (such as looking for donors in the census imputation system) is based on whole local authority areas within design groups, while GROS uses health Board Area based design groups after matching within the One Number census system. Therefore the donor can only be sought from within the Bailery, while ONS could stretch the search to the Design Group.

### **2.1 Sampling Units**

2.11 The CCS is a postcode-based survey. A sample of postcodes, rather than households, will be re-enumerated.

2.12 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 employs a two-stage cluster design with 1991 EDs (In Scotland, actually 1991 Output Areas mapped back to 1981 EDs<sup>1</sup>) as primary sampling units (PSUs) and postcodes within EDs as secondary sampling units (SSUs). The following section describes the stratification of the PSUs.

---

<sup>1</sup> GROS followed the ONS sample design as closely as possible. ONS used 1991 EDs as the Primary Sample Units because their output data was aggregated by enumeration district. However, GROS had a separate output geography based on aggregations of postcodes. The Output Areas were about 1/5<sup>th</sup> the size of the EDs so they were too small to be used as primary Sample Units. Therefore 'pseudo' 1981 EDs were used.

**Table 1 The population and Council Areas of the 8 Design Groups in Scotland.**

<b>Design Group 'name'</b>	<b>Health Board Areas</b>	<b>Pop.</b>	<b>Council and part Council Areas</b>	<b>Population</b>	<b>Council Area</b>
DAGAAB 629,000	Dumfries & Galloway	147,300	Dumfries and Galloway	147,300	Dumfries & Galloway
	Ayrshire & Arran	375,400	North Ayrshire	139,660	North Ayrshire
			South Ayrshire	114,440	South Ayrshire
			East Ayrshire	121,300	East Ayrshire
Borders	106,300	Scottish Borders	106,300	Scottish Borders	
LOTHIAN 773,700	Lothian	773,700	East Lothian	89,570	East Lothian
			Midlothian	80,860	Midlothian
			City of Edinburgh	450,180	City of Edinburgh
			West Lothian	153,090	West Lothian
FOVAFIFE 624,700	Forth Valley	275,800	Falkirk	144,110	Falkirk
			Stirling	83,130	Stirling
			Clackmannanshire	48,560	Clackmannanshire
	Fife	348,900	Fife	348,900	Fife
GRAMPIAN 525,200	Grampian	525,200	Aberdeen City	213,070	Aberdeen City
			Aberdeenshire	226,260	Aberdeenshire
			Moray	85,870	Moray
TAYHOSE 668,500	Tayside	389,800	Dundee City	146,690	Dundee City
			Angus	110,070	Angus
			Perth & Kinross	133,040	Perth & Kinross
	Highland	208,300	Highland	208,300	Highland
	Orkney	19,550	Orkney	19,550	Orkney Islands
	Shetland	22,910	Shetland	22,910	Shetland Islands
	Western Isles	27,940	Western Isles	27,940	Eilean Siar
CLYDECUMB 560,800	Lanarkshire	560,800	Clydesdale, East Kilbride and Hamilton Districts	250,300	South Lanarkshire
			Cumbernauld & Kilsyth, Monklands and Motherwell Districts	310,500	North Lanarkshire
GREATER GLASGOW 911,200	Greater Glasgow	911,200	Rutherglen	56,560	South Lanarkshire
			Chryston (NL1)*	16,220	North Lanarkshire
			Glasgow city	619,680	Glasgow City
			Eastwood	63,050	East Renfrewshire
			East Dunbartonshire	109,570	East Dunbartonshire
			Clydebank	46,120	West Dunbartonshire
ARGYLL & CLYDE 426,900	Argyll & Clyde	426,900	Barrhead part area	24,930	East Renfrewshire
			Dumbarton District – Helensburgh part area (27,320)	48,760	West Dunbartonshire
			Inverclyde	85,400	Inverclyde
			Renfrewshire	177,830	Renfrewshire
			Argyll & Bute	89,980	Argyll & Bute
<b>Total</b>				<b>5,120,000</b>	

## 2.2 Stratification of PSUs using the Hard-to-Count Index.

2.21 It is expected that under enumeration in the 2001 Census will be higher in areas characterised by particular social, economic and demographic characteristics. For example, one would expect people in dwellings occupied by more than one household (multi-occupancy) to have a relatively high probability of not being enumerated in a census. In order to control for this, EDs within each Design Group are stratified by a Scotland 'Hard to Count' (HtC) score. This score is calculated by combining characteristics that were found to be important determinants of under enumeration by the Office for National Statistics (ONS) and the Estimating With Confidence Project (Simpson *et al.*, 1997).

2.22 The HtC score for 2001 in Scotland is that used in the 1999 Rehearsal. It is formed using the following variables from the 1991 Census:

- a) Households in multi-occupied buildings
- b) Young Migrant households
- c) Households where head is non-white
- d) Imputed households
- e) Households living in Private Rented Accommodation

2.23 ONS will use the HtCI for sample stratification and workload weighting for interviewers. Therefore, the derivation of the HtCI waited for both the output from the 1999 rehearsal, which in the event was late, and from an analysis of the field staff data (e.g. hours worked against HtCI) in England and Wales.

2.24 The rehearsal data from Scotland did not show a relationship between the 1999 HtCI and the time taken by an enumerator or the results achieved. In fact, the time taken by an interviewer, on a general level, was determined overwhelmingly by the rurality of a postcode. There were also very specific, local factors that determined the hours worked by an interviewer, e.g. new blocks of student flats (which could not affect the 1991 based HtCI). Therefore for 2001, postcode size is used to determine workloads in rural areas only, and all urban areas have about 95 households in them. A flexible support system was introduced to deal with unexpected problems within work loads.

2.25 The HtCI was not used to determine the workloads in Scotland. This weakened one need for a rigorous HtCI. In Scotland, the HtCI is therefore only used as stratification for the sampling and estimation.

2.26 GROS followed a draft proposal of ONS (Brown *et al* 1999) but altered the ethnic factor. The GROS HtCI uses a simple White/non-White ethnic factor, because, in many design groups, a large minority, if not majority, of people born in non-English speaking countries were of white ethnic origin (see Table 2). This would have meant that, had we used 'proportion of persons whose country of birth is non-English speaking' (the variable proposed by ONS), the distribution of HtCI group 3 would have responded to the distribution of white settlers.

2.27 ONS proposed an HtCI to the ONC Steering Committee in June 2000 (ONC(SC)00/15). It included an unemployed factor. The Committee agreed that there was a relationship between under enumeration and unemployment despite an earlier view that the unemployed may be easier to

contact in a coverage survey. The inclusion of this factor is probably an improvement on the GROS stratification, but by the time the proposal was agreed, GROS had already taken its sample

Table 2: Country of Birth by Ethnic Group in Scotland 1991.

	White	Black-Caribbean	Black-African	Indian	Pakistani	Bangla deshi	Chinese	Others
UK + Eire	4,844,754	495	711	4,624	11,509	417	3,286	7,868
AUSCANNZ	15,192	5	12	10	2	1	13	125
Africa	4,742	9	1,375	596	133	3	1	435
Caribbean	996	306	9	45	2	0	38	234
Indian Sub-continent	4,300	7	15	4,482	9,452	701	22	928
Asia	4,294	4	6	115	28	1	5,312	1,011
Islands	3,573	4	41	35	10	0	140	369
Europe	34,329	4	21	10	24	0	39	689
N Africa	1,072	3	77	1	2	0	2	774
Other Africa	4,547	2	301	50	3	0	5	154
USA	11,843	9	85	28	0	0	9	615
Other Caribbean	234	11	1	0	0	0	2	100
S America	1,461	6	0	2	1	0	2	256
Arabia etc	2,106	2	8	16	24	11	4	1,784
Asia other	1,373	1	1	24	2	0	1,601	1,925
Other	72	0	0	12	0	0	0	29
	4,934,888	868	2,663	10,050	21,192	1,134	10,476	17,296

2.28 It was not worthwhile re-sampling because :

- The HtCI is based on data from 10 years ago;
- Unemployment is relatively low;
- It was unlikely that an unemployment factor added much as an explanatory variable to the other five variables.
- The main purpose of the HtCI was as a stratification to limit the random nature of the sample. It also ensures that the CCS sample is spread across the full range of EDs. This is achieved with the 5 factors selected by GROS.

2.29 As in ONS, the HtCI score is derived by a summation of the proportions of the HtCI.

### 2.3 Overall Design

2.31 The second stage of the CCS design consists of the random selection of postcodes within each selected primary sampling unit. ONS investigated the number of postcodes to be chosen within each ED in paper ONS(ONC(SC))98/12. Their research indicated that a maximum of five postcodes per ED would provide a statistically efficient design and a relatively efficient allocation of resources while still maintaining a robust approach. On average, a postcode has about 15 households. Therefore, on average, ONS will have selected 75 households in 5 postcodes at a sample spot.

2.32 GROS adapted the ONS approach. For an efficient as possible allocation of resources, given the geographical spread of the sample in Scotland, a number of whole postcodes were selected from the PSU until about 100 households (based on a modified delivery point count from the Postal Address File) were sampled. A single interviewer can cover 100 households in about 80 hours

within a small geographical area. Therefore, at each PSU a sample equal to 1 interviewer workload was chosen. In rural areas, fewer households were selected because of the travel involved.

## 2.4 Sample size and distribution

2.41 To achieve the aims of the CCS, the sample size must be sufficiently large to enable population estimates of an acceptable degree of precision. The ONS simulation of their design (using a number of simulated 1991 Censuses and Surveys) indicated that the optimal sample size representing the best value for money in terms of precision was 20,000 postcodes for England and Wales (for a population of about 52,000,000), implying 4,000 PSUs. This research is presented in ONS(ONC(SC))98/12. This is a sample size of about 1.4%.

2.42 GROS had limited resources to repeat the ONS simulation, though some simulations were carried out to check that the final sample gave modelled accuracy levels similar to ONS. Also, sampling was carried out very early in the survey project development (June 2000) because of the need to re-deploy staff to other Census and CCS activities. There was always the danger that having taken an early sample, ONS would change the sample size. Therefore, a proportionately larger sample was taken to guard against such a danger. The sample size chosen was therefore 400 PSUs for a population of approximately 5,000,000. However, a further 10 PSUs were purposively selected to ensure some coverage in all 36 Bailerries.

2.43 As described above, on average more than 5 postcodes were selected from each PSU. Also, in rural areas especially, postcodes often have far fewer than 15 households and so some sample points had up to 26 postcodes. This also happened in city shopping and commercial areas with few residents. The final GROS sample contains 2,377 postcodes. In terms of households, the CCS sample is estimated to include 39,644 households. Given an average household size of 2.3 people, the sample size is about 1.8%.

2.44 It is expected that the under enumeration in 2001 will not be evenly spread across the country. Therefore, it is sensible to weight the sample towards the areas that are expected to have a higher undercount. The aim is to produce Design Group estimates with comparable accuracy. Therefore, the amount of the sample allocated to each strata must be that which gives a similar expected precision. The actual obtained precision will be dependent on the population size of the Design Group, the level of under enumeration in the 2001 Census and the CCS sample size. Within each Design Group, it is expected that the variance will be higher in the hardest to count EDs and hence there will be a larger sample size in the hardest to count areas.

2.45 The final allocation of the PSUs was about:

- 3.0% of HtC category 3 EDs (the hardest);
- 2.5% of HTC category 2; and
- 2.0% of HTC category 1 (the easiest)<sup>2</sup>.

Therefore, a Design Group made up of mostly hard to count EDs should be allocated a larger sample size than a Design Group made up of easy to count areas. The relative sample rates<sup>3</sup> by HTCI for each Design Group and their component Council Areas are shown in Table 3.

---

<sup>2</sup> The PSU sample rate is higher than the national person sample rate because only a proportion of the PSU was sampled.

<sup>3</sup> The relative sample rate is the number of PSUs multiplied x the average number of households at a sample point (95) \* the average number of people in a household (2.3) divided by the population times the average percentage of PSUs in a Hard to

**Table 3: PSUs and Relative Sample Rates by HTC Group, Bailery and Design Group**

Bailery grouped by Design Group	PSUs in each HTC Group			Total EDs (PSUs)	Population	Relative Sample Rates (%)			Total %
	1	2	3			1	2	3	
Dumfries and Galloway	4	7	1	12	147,300	1.5	2.6	0.7	1.8
East Ayrshire	5	5	2	12	121,300	2.3	2.3	1.8	2.2
North Ayrshire	3	1	2	6	139,660	1.2	0.4	1.6	0.9
Scottish Borders	2	4	4	10	106,300	1.0	2.1	4.1	2.1
South Ayrshire	6	3	2	11	114,440	2.9	1.4	1.9	2.1
	<b>20</b>	<b>20</b>	<b>11</b>	<b>51</b>	<b>629,000</b>	<b>1.7</b>	<b>1.7</b>	<b>1.9</b>	<b>1.8</b>
City of Edinburgh	8	11	15	34	450,180	1.0	1.3	3.6	1.7
East Lothian	2	2	4	8	8,219	13.3	13.3	53.2	21.3
Midlothian	2	4	1	7	80,860	1.4	2.7	1.4	1.9
West Lothian	6	7	1	14	153,090	2.1	2.5	0.7	2.0
	<b>18</b>	<b>24</b>	<b>21</b>	<b>63</b>	<b>692,349</b>	<b>1.4</b>	<b>1.9</b>	<b>3.3</b>	<b>2.0</b>
Clackmannanshire	1	1	1	3	48,560	1.1	1.1	2.2	1.3
Falkirk	10	2	1	13	144,110	3.8	0.8	0.8	2.0
Fife	8	13	9	30	348,900	1.3	2.0	2.8	1.9
Stirling	2	5	3	10	83,130	1.3	3.3	3.9	2.6
	<b>21</b>	<b>21</b>	<b>14</b>	<b>56</b>	<b>624,700</b>	<b>1.8</b>	<b>1.8</b>	<b>2.4</b>	<b>2.0</b>
Aberdeen City	6	6	4	16	213,070	1.5	1.5	2.1	1.6
Aberdeenshire	4	12	7	23	226,260	1.0	2.9	3.4	2.2
Moray	1	2	3	6	85,870	0.6	1.3	3.8	1.5
	<b>11</b>	<b>20</b>	<b>14</b>	<b>45</b>	<b>525,200</b>	<b>1.1</b>	<b>2.1</b>	<b>2.9</b>	<b>1.9</b>
Angus	1	2	5	8	110,070	0.5	1.0	5.0	1.6
Dundee City	6	9	4	19	146,690	2.2	3.4	3.0	2.8
Highland	2	9	4	15	208,300	0.5	2.4	2.1	1.6
Orkney	0	3	1	4	1,219	0.0	134.5	89.7	71.7
Perth & Kinross	1	3	3	7	133,040	0.4	1.2	2.5	1.1
Shetland	1	1	1	3	22,910	2.4	2.4	4.8	2.9
Western Isles	2	1	1	4	27,940	3.9	2.0	3.9	3.1
	<b>13</b>	<b>28</b>	<b>19</b>	<b>60</b>	<b>650,169</b>	<b>1.1</b>	<b>2.4</b>	<b>3.2</b>	<b>2.0</b>
Clydesdale SL2	9	10	2	21	250,300	2.0	2.2	0.9	1.8
Cumbernauld NL2	13	9	4	26	310,500	2.3	1.6	1.4	1.8
	<b>22</b>	<b>19</b>	<b>6</b>	<b>47</b>	<b>560,800</b>	<b>2.1</b>	<b>1.9</b>	<b>1.2</b>	<b>1.8</b>
Chryston NL1	2	1	2	5	16,220	6.7	3.4	13.5	6.7
Rutherglen SL1	2	3	1	6	56,560	1.9	2.9	1.9	2.3
Glasgow city	12	22	11	45	619,680	1.1	1.9	1.9	1.6
East Dunbartonshire	8	3	1	12	10,219	42.8	16.0	10.7	25.7
Eastwood ER2	2	4	1	7	63,050	1.7	3.5	1.7	2.4
Clydebank WD2	6	3	0	9	46,120	7.1	3.6	0.0	4.3
	<b>32</b>	<b>36</b>	<b>16</b>	<b>84</b>	<b>811,849</b>	<b>2.2</b>	<b>2.4</b>	<b>2.2</b>	<b>2.3</b>
Argyll & Bute	1	4	5	10	89,980	0.6	2.4	6.1	2.4
Barrhead part ER1	2	1	0	3	24,930	4.4	2.2	0.0	2.6
Dumbarton WD1	0	5	2	7	48,760	0.0	5.6	4.5	3.1
Inverclyde	5	3	3	11	85,400	3.2	1.9	3.8	2.8
Renfrewshire	6	2	1	9	177,830	1.8	0.6	0.6	1.1
	<b>14</b>	<b>15</b>	<b>11</b>	<b>40</b>	<b>426,900</b>	<b>1.8</b>	<b>1.9</b>	<b>2.8</b>	<b>2.0</b>
<b>Total</b>	<b>151</b>	<b>183</b>	<b>112</b>	<b>446</b>	<b>4,920,966</b>	<b>1.7</b>	<b>2.0</b>	<b>2.5</b>	<b>2.0</b>

Count Group. Therefore if the population of a design group is 500,000, 20% should be in HTCI =3 or 100,000. If I have 3 PSUs of HTCI = 4 then the sample size is  $9 * 95 * 2.3 = 1,966$  people and a nominal relative sample of 1.96%.

### **3. Matching the Census Coverage Survey and Census Records**

3.1 There is no difference in the matching strategy or detail between GROS and ONS. GROS may carry out the manual matching at GROS if this proves feasible, but is committed to ensuring the prescribed standards of the project are matched. GROS are currently developing and testing a telecommunications link to ONS for matching at GROS. Failing that, GROS will have staff in place in Titchfield for the matching, including a Gaelic speaker.

### **4. Estimation of Design Group and Council Area Age-Sex Populations**

4.1 There is no significant difference in the estimation strategy or detail between ONS and GROS. ONS will output and pass the data for the 36 Bailerries to GROS for quality assuring.

### **5 Small area estimation and imputation**

5.1 There is no significant difference in the estimation and imputation strategy or detail between ONS and GROS.

### **6. Demographic Estimates and Quality Assurance**

6.1 While the 2001 Census-based, ONC estimates will be considered as the 'Standard', there may be practical problems in the CCS so it is important that there is a quality assurance process. This QA process is under development following the strategy laid out in a Scottish Census Advisory Group Paper (SCAG 00/21). Central to the QA process is the use of the best possible comparable demographic estimates as well as data from other administrative sources that can serve as an independent check on the plausibility of ONC estimates.

6.2 Scotland has only 8 Estimation Area, 15 Health Boards, 32 Council Areas and 36 Bailerries. Except for the first Estimation Area, the Estimation Areas will all be processed consecutively and early in the Census downstream processing timetable. The current plan is therefore to quality assure all the estimations and imputation totals as a single operation.

6.3 Detailed QA and contingency proposals are being drawn up by GROS for use in operational trials.

## 7. References

Brown, J., Diamond, I., Chambers, R., Buckner, L. and Teague, A (1999) A methodological strategy for a one-number census in the UK. *J. R. Statist. Soc. A* 162 (2), 247-267.

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

ONS(ONC(SC))00/15 - 2001 Hard to Count Index. Titchfield: ONC

Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends*, 90, 31-39.

SCAG 00/21, 2000. Quality Assurance and Contingency Strategy for the ONC – Scotland. GROS.