



ONS(ONC(SC))99/01

ONE NUMBER CENSUS STEERING COMMITTEE

One Number Census Project Update

1. This paper describes the progress on the One Number Census since the last meeting of the ONC Steering Committee on 13 November 1998. It also outlines the work planned for the next 6 – 9 months.
2. **The Steering Committee are asked to:**
 - a) **note the progress made and endorse the research programme**
 - b) **provide any comments at the meeting on the 1 July 1999, or in writing by 15 July 1999.**

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

June 1999

One Number Census Project Update

Marie Cruddas

1. Introduction

1.1 This paper reports progress on the One Number Census (ONC) project since the last meeting of the ONC Steering Committee on 13 November 1998. Work planned for the next 6-9 months is also included.

1.2 The work into implementation and development of the Census Coverage Survey (CCS) in 2001 is covered by a separate CCS Project under the management of John Dixie. ONS(ONC(SC))99/02 provides an update of progress on the CCS project since the last meeting of the Steering Committee.

2. Update on key areas of research

Census Coverage Survey Design

Hard to Count index

2.1 The Hard to Count (HtC) index has been simplified for the 1999 Census Coverage Survey Rehearsal. Three variables are being used: Multi-occupancy, Private Rented accommodation and households containing young migrants. The percentages of these variables are summed to form the score, which is still at present divided into five equal sized groups at the national level. This HtC index will be evaluated following the Census Rehearsal and further research will be undertaken to examine the variables included in the index and the number of categories.

1999 Census Coverage Survey Rehearsal Sample

2.2 The sample for the 1999 Census Coverage Survey Rehearsal in England and Wales was selected in January 1999. This used the revised HtC index described above. The Census Rehearsal areas have a very skewed distribution of HtC categories – the majority of areas being the hardest to count. As a result, a constraint was put on the sampling to ensure that every HtC group in each Rehearsal area was represented in the sample. This was to ensure that sufficient data are available to evaluate the index.

2.3 It was found that the initial selection of 1200 postcodes contained far more addresses than was expected (approximately 26,000 according to Royal Mail data). The primary reason for this was that the postcodes in the Rehearsal areas had a higher mean number of addresses (23) than the overall national mean (15). This was a consequence of the areas selected for the Census Rehearsal which were not nationally representative but chosen to address the enumeration of certain types of people and accommodation e.g. students, ethnic minority groups, small hotels, etc.

2.4 Therefore, the selection process was carried out a second time to reduce the number of expected addresses. This resulted in a sample size of 818 postcodes with an expected 18,000 addresses – which equates to a 17% sample of postcodes in the Census Rehearsal areas.

2.5 In Scotland the sample size was 130 postcodes in Dundee and Angus which is approximately 1800 addresses and the CCS Rehearsal in Northern Ireland covered approximately 550 addresses.

Design Groups for 2001

2.6 An initial grouping of Local Authority Districts (LADs) and Unitary Authorities (UAs) into Design Groups has been made for England and Wales. The criteria used for constructing Design Groups/Estimation Areas are listed below.

1. Design Groups MUST respect the Welsh and Scottish borders.
2. Design Groups must either be whole LADs/UAs or groups of whole LADs/UAs (no LADs/UAs will be split across Design Groups).
3. If possible, LADs/UAs in the Design Group must be contiguous (or ‘close’ together, i.e. in the same area of the country - North East, South West).
4. Each Design Group will have a total population of approximately 500,000 – large LADs/UAs with a population greater than 500,000 people will not be split for ONC purposes.

Where possible:

5. The number of LADs/UAs in a Design Group should be minimised (less than 8 if possible). If there are a large number of LADs/UAs then the possible sampling constraint that at least one Enumeration District (ED) from each HtC group in each LAD/UA should be selected may result in the selection being too restrictive and a larger than expected sample size.
6. The expected characteristics of the undercount in LADs/UAs in a Design Group should not be too dissimilar (HtC 5 is same across all LADs in group).
7. LADs/UAs of a similar size should be grouped
8. County borders will be respected

2.7 For Scotland a similar set of criteria have been used, namely that each design group is a collection of contiguous Council areas (except for Edinburgh and Glasgow which each form a single design group and Aberdeen and Dundee which together form a design group but are not contiguous). At present an initial grouping has produced 11 design groups each of which groups council areas which are similar in character subjectively - i.e. predominantly rural or predominantly industrial. It is intended that this number will be reduced to 10 design groups. After this reduction the Scottish Local Authorities will be consulted about the proposed grouping for their comments.

2.8 In Northern Ireland, it is likely that there will be three design groups each consisting of combinations of Local Government districts.

Matching

2.9 This section describes the progress made in the ONC matching methodology since the work outlined in the paper ONS(ONC(SC))98/14. The key areas covered here are:

- Matching weights
- Household classifications
- Matching software

Matching Weights

2.10 Probability matching involves assigning a probability weight to a pair of records based on the level of agreement between the two records. The probability weights reflect the likelihood that the two records correspond to the same individual. Two records may then be assigned as a match, even if they disagree on a small number of details, provided the probability weight exceeds a pre-determined threshold.

2.11 Before probability matching can be undertaken, it is necessary for initial probability weights to be in place. These weights do not need to be precise for the matching to be of a reasonable quality. Provided the starting weights are close to the appropriate values, they can be fine-tuned during the matching process.

2.12 Therefore, it is proposed to use the weights derived from the 1999 Census Rehearsal as starting weights for the 2001 matching process. These weights will then be updated as the matching process continues to ensure that they are appropriate to the data being matched. This process will remove the necessity for a time-consuming clerical match to be performed prior to the automated matching of the 2001 CCS and Census data.

2.13 The precise methodology for updating these weights will be produced as part of the evaluation of the Census Rehearsal.

A Household Classification

2.14 A household classification can be used to combine dependent variables, such as the number of people in a household, their ages and relationships, into a single matching variable providing good distinguishing information.

2.15 The proposed household classification partitions households into the 20 categories given in Table 1. This classification includes rules for missing values as it is possible that it will be applied to the data before item imputation has been performed.

2.16 The household classification has been applied to the October 1998 CCS test data. The number and percentage of households falling within each category are also shown in Table 1. The classification will be reviewed as part of the Census Rehearsal evaluation.

Matching Software

2.17 The prototype matching system will be programmed in Visual C++. Work currently being undertaken implies that Visual C++ is a suitable programming language with which to

implement the system, and interfaces with the Census Database and form images have been established. The prototype matching system will be evaluated as part of the Census Rehearsal.

2.18 A meeting was held with Dr. Steve Kendrick the ONC matching consultant. He approved the approach for the Census Rehearsal and is keen to be more involved when the matching is in progress using Census Rehearsal data.

<i>HOUSEHOLD TYPE</i>	<i>Category</i>	<i>Number of Households</i>	<i>Percentage</i>
Single Person Household			
• Person aged over 65	11	186	10%
• Person aged under 40	12	136	8%
• Other (including missing age)	13	146	8%
Two Person Household			
• Couple, both over 55	21	233	13%
• Couple, one or both under 56	22	211	12%
• Other related	23	74	4%
• Other (including missing relationship, couples with a missing age)	24	66	4%
Three Person Household			
• Couple and child aged under 16 (including step-child)	31	73	4%
• Couple and child aged 16 or over (including step-child)	32	73	4%
• Other related	33	56	3%
• Other (any relationship missing or unrelated or child's age missing)	34	48	3%
Four Person Household			
• Couple & two children, at least one aged less than 16 (including step-children)	41	105	6%
• Couple & two children, both aged 16 or over (including step-children)	42	37	2%
• Other related	43	37	2%
• Other (including any relationship missing, any person unrelated, any child's age missing)	44	71	4%
Five Person Household			
• Couple and three children	51	67	4%
• Other	52	56	3%
Six Person Household	61	52	3%
Seven+ Person Household	71	40	2%
Other household (i.e. no individual details)	81	20	1%
<i>TOTAL</i>		<i>1,787</i>	

Table 1: Table showing the classification of households into 20 categories and how these categories partition the 1998 CCS test data.

Design Group and Local Authority District Estimation

Design Group Estimation

2.19 The paper ONS(ONC(SC))99/07 sets out to investigate the use and robustness of the joint regression DSE approach for general population estimation using simulations. In particular it considers:

- a) The robustness of the DSE to heterogeneity and dependence.
- b) The robustness of the regression model compared to the simpler ratio model.
- c) Unbiased estimation from regression estimation.
- d) The level at which the DSE is calculated.
- e) Calculation of population totals for variables other than age and sex.

2.20 Results from the paper demonstrate that the DSE part appears reasonably robust while the regression estimator has some problems. It turns out that the DSE with a ratio model is 'better' in general and there is little to choose between defining the DSE at the postcode level or the HtC level. However, it is clear that the 'optimal' strategy will involve mixing regression and ratio estimation and what still needs to be answered is how to define the 'optimal' approach for a given set of data. The paper also demonstrates through simulations a simple theoretical result that occurs when there is matching error. This is early work but it shows that errors at the matching stage induce bias in the DSE and population estimation.

2.21 Initial work using a more complex model to cope with the large number of times that both the census and CCS find no people, as well as the few cases where only the CCS finds people in the postcode, has as yet not produced the expected reduction in bias. For this reason there are no results in the paper. However, this is ongoing research and relates to choice of which strategy to use in a particular situation.

LAD estimation

2.22 Simulation work has demonstrated that estimates can be made at LAD level with a good degree of precision using a simple synthetic estimator. The results from this were used to form the expected LAD precision tables in 'A Guide to the One Number Census'.

2.23 Work is underway to establish the different possible approaches to the One Number Census LAD estimation problem and to examine the rationale, assumptions, advantages and disadvantages of each.

2.24 It is expected that an evaluation of the different approaches will not result in a 'best method' for all circumstances. The outcome is likely to be a defined strategy that determines which approach to use under particular conditions, or when some predefined criteria are met.

2.25 The evaluation of the approaches will involve two stages. The first will be to explore each approach by fitting the models to a typical dataset, and refining if necessary. Once this has been completed, large scale simulation studies will attempt to examine the relative performance and robustness of the approaches under different conditions.

Demographic Analyses and Administrative Records in Support of a One Number Census

2.26 Since the last Steering Committee meeting preliminary uncertainty intervals have been constructed for 2001 national level demographic estimates. Paper ONS(ONC(SC))99/05 covers this in detail.

2.27 For the One Number Census process some demographic estimates will be required at the subnational level, that is for the design groups for which estimates are produced directly from the Census Coverage Survey. The precise requirements of the One Number Census for demographic estimates at the subnational level are still to be specified. Demographic subnational estimates can be based on the rolled forward population estimates, as the design groups are straight aggregations of the local authority areas for which these population estimates are produced.

2.28 Administrative registers will be used at the national and subnational level to provide population counts for some population subgroups. Research work is being planned to develop the use that will be made of the administrative registers. Previous work concluded that no register could be found which was suitable for use at an individual level but identified sources that showed potential for use at an aggregate level. Since the last Steering Committee meeting, there has been agreement for us to use the following for 1999 Census Rehearsal:

- Health registration data - FHSA
- DVLA
- DSS Pensions and Child Benefit
- Student data - HESA

Work has now begun to obtain and assess the quality of these data and also to derive a strategy for their use in the 1999 Census Rehearsal. This work will be carried out in liaison with the Census Data Quality project and Population and Vital Statistics.

Imputation and weighting

2.29 Since the last Steering Committee much of the work has focused on improving the implementation of the methodology on the computer. There has been a particular focus on reducing pruning and grafting as this is the process that is least controlled. To reduce pruning and grafting:

- a) The model for household coverage weights has been extended to include characteristics of individuals within the household. This ensures that the imputed households contain the right sort of people.
- b) There is an intermediate calibration of the individual coverage weights to reflect the people imputed at the first stage.

These two things have stopped the over imputation of the 45-79 year olds, as the household imputation no longer imputes so many households containing these people.

2.30 The way the data is processed has also been looked at. Impute classes are now purely defined by the coverage weights. However, these still discriminate based on the characteristics that are used to model the weights including tenure for households and age-sex for individuals. There is nothing to stop other variables also being included to define impute classes but this is a refinement that can be considered at a later date. The definitions of variables in the datasets for modelling have also been changed. Variables such as size, type and marital status are defined using the combined census-CCS data. This makes the donor and recipient searches easier and prevents the problem of looking for impossible people.

2.31 The result of the changes has been a major improvement in efficiency with respect to computing time and the whole process is now down to about 48 hours of processor time for a design group. Gains can still be made through further optimisation of the SAS programs.

2.32 A short paper summarising the methodology with a few results has been produced for the Statistics Canada Symposium '99 Proceedings. This has been circulated to Steering Committee members for comment - ONS(ONC(SC))99/08

2.33 The ONC Imputation SAS code has now been handed over by Southampton University to ONS for integration into the ONC prototype system for the Census Rehearsal, further development and optimisation.

3. Update on Implementation

3.1 Work is continuing to develop a prototype ONC system for use with the Rehearsal data. This will be used to evaluate the methodology developed for the ONC. The next significant challenge for the project is to integrate the ONC into the 2001 Census programme in an effective way. To do this we need to think carefully through the strategic options and clarify the requirements before committing to any development of systems.

3.2 One of the key issues in implementing the ONC will be to ensure that it can be carried out to meet key targets. The processing timetable for 2001 is discussed further in ONS(ONC(SC))99/03.

3.3 A document outlining the ONC System Requirements has been drafted. It is being updated by inclusion of a discussion of the strategic options for the development and running of the systems.

3.4 As part of the testing and evaluation of the ONC prototype system 1991 Census data will be used to simulate a census and CCS. These data will then be run through the system which will allow the following to be investigated prior to the availability of Census Rehearsal data:

- Interfaces between the Census database and the ONC systems.
- Interfaces between different parts of ONC system, e.g. matching, estimation and imputation.
- Obtain a broad based estimate of how long each system will take to run and how many people may be required to manage it for a design group. This will not be possible for the prototype

matching system which will require Census Rehearsal data to determine estimates of the time and resources due to the differences in questions between 1991 and 2001 and the difficulties with simulating appropriate errors in the data.

Scotland and Northern Ireland

3.5 Discussions have been held between ONC representatives in ONS, GRO(S) and NISRA and it has been agreed that for the Census Rehearsal all ONC processing will be undertaken at Titchfield with GRO(S)/NIRSA providing additional staff for short periods if required. GRO(S) and NISRA representatives will also be welcomed and encouraged to work closely with the ONS ONC team at the time of the ONC Rehearsal processing to ensure that any additional or different requirements that they have are met. Any estimates produced by the process for Scotland and Northern Ireland will be passed to GRO(S)/NISRA for testing of the QA procedure and agreement of control totals for the imputation system.

4. Programme of further work

4.1 The majority of the work planned for the next 6 – 9 months will be driven by the further development of prototype systems and subsequent evaluation of the Census Rehearsal. This evaluation is covered separately in ONS(ONC(SC))99/04.

CCS design

4.2 Further research will examine whether the 1991 Census based variables used in the HtC index for the Census Rehearsal are sufficiently suitable for small areas, using 1997 Test and 1999 Rehearsal data. This, and other aspects, will be included within a detailed evaluation of the index following the 1999 Rehearsal (ONS(ONC(SC))99/04). Further work is planned to examine the number of levels the index should have and the methodology for assigning the HtC categories.

4.3 A number of issues have already been identified through the 1999 sampling process, the main issue being the quality of the data used to form the sampling frame. This will be examined in more detail in the evaluation of the sampling for the Rehearsal although research is underway to assess the options.

Finalise Design Groups for 2001.

Matching

- Select matching and blocking variables, in particular evaluating the contribution of individual's names and addresses to the matching process
- Methodology for tuning the matching weights to be developed
- Methodology for estimating false match rates to be developed
- Produce prototype systems to assist the clerical matching and perform the automated matching of the Rehearsal data.

Design Group and LAD Estimation

4.4 This will mainly concentrate on investigating the theoretical properties of the estimators. This work, backed-up by some further simulations should give insight into the following:

- a) The most stable level for doing the DSE. The three choices are postcode, cluster, and HtC index.
- b) Simple rules that the computer can apply for choosing between the ratio and regression models. It should be noted that while a Horvitz-Thompson type estimator has been rejected as a general approach (it does not make efficient use of all the 2001 Census data) it may be more suitable in specific situations.
- c) Whether a more complex model to explicitly deal with the zero counts in the census can lead to unbiased estimation.

4.5 There is also the outstanding question of variance estimation. The 'Ultimate Cluster Variance Estimator' was used in the early simulations and gives good coverage. However, there are other alternatives to be considered.

Imputation and weighting

- Complete simulations and fully assessed how well the process works for any particular simulation as well as its stability across simulations.
- Produce a detailed paper of the methodology with full results for publication in a refereed journal
- Investigate the introduction of dummy forms into the simulations. At present there is little evidence to suggest how to create them. (Creating a perfect set of dummy forms in the simulation is simple but this probably doesn't reflect the quality that could be expected in the 2001 Census.) However, this is an important area that will need to be investigated and it is expected that their use will improve the placement of imputed households across the EDs.