



ONS(ONC(SC))00/19

ONE NUMBER CENSUS STEERING COMMITTEE

1999 Census Rehearsal: ONC Evaluation Report

1. This paper reports progress on the evaluation of the 1999 Census Rehearsal of the ONC process.
2. **Members of the Steering Committee are asked to note progress on the evaluation.**

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

June 2000

1. INTRODUCTION

This document summarises progress against the planned evaluation of the One Number Census (ONC) processes described in the ONS(ONC(SC))00/06 1999 CENSUS REHEARSAL: ONC EVALUATION PLAN.

The key areas of the ONC evaluation are:

- CCS sample selection;
- Matching the Rehearsal CCS and Census data;
- Estimation;
- Demographic estimates and administrative records;
- Imputation;
- ONC Overview.

Progress is reported against each of the evaluation objectives contained in the evaluation plan.

As a consequence of the late delivery of the rehearsal data it has not been possible to perform all of the planned evaluation. This impacts in particular on evaluation of the ONC Matching process where we have by necessity had to concentrate on the quality and speed of clerical matching. For other aspects of the ONC it has been possible to progress the evaluation by using simulated data.

2. CCS SAMPLE SELECTION

2.1 Background

The CCS involves an intensive enumeration of a large, nationally representative sample of postcodes. The sample of postcodes will be drawn from all Local Authority Districts (LADs) to enable population estimates to be made for all districts. LADs will be grouped together to form 'design groups' which have a population of approximately half a million people. The CCS sample design is applied to each design group independently. The strategy for the selection of postcodes within each design group is given in papers ONS(ONC(SC))00/01 and ONS(ONC(SC))98/12. The design groups for 2001 are presented in ONS(ONC(SC))00/10.

2.2 Evaluation Topics

The evaluation of the sample selection focused on five areas:

- The sampling frame;
- The Hard to Count index;
- Size stratification;
- The prototype sampling system;
- Timing.

2.2.1 *Sampling Frame*

Objective: To evaluate the quality of the CCS sampling frame and determine an appropriate sampling frame for 2001.

Outcome: The data sources and methodology used for the sampling frame were both appropriate for the rehearsal. A number of issues were identified:

- A Hard to Count index is required for all 1991 EDs, including those that failed the 1991 thresholding for output;
- More up to date Postcode Address File extracts should be used for the selection of postcodes;
- Each 1991 ED should have a minimum of 1 postcode linked to it;
- The selection of the postcodes should occur as late as possible to minimise the risk of changes; and;
- The postcode selection should be monitored against the latest Postcode information.

2.2.2 Hard to Count Index

Objective: To evaluate the Hard to Count index and determine whether the variables used are appropriate for inclusion in the 2001 index.

Outcome: The evaluation work carried out is contained in Paper ONS(ONC(SC))00/15

2.2.3 Size Stratification

Objective: To evaluate the data used to implement the CCS design size stratification that occurs within each Hard to Count category.

Outcome: The data used is appropriate and therefore no synthetic adjustment is necessary.

2.2.4 Prototype Sampling System

Objective: To evaluate the prototype sampling system and determine the most appropriate language for implementing the 2001 system.

Outcomes:

- The system is sufficiently easy to use;
- The prototype is suitable as the basis for the 2001 system;
- A GIS function should be added to the system to allow the examination of the geographical distribution of the sample;
- The 2001 Sampling system should be implemented using SAS.

2.2.5 Timing

Objective: To determine the time required to perform the sample selection for 2001.

Outcome: It is estimated that it will take 1 person approximately 3 weeks to complete the selection of the sample for all 2001 Design Groups. The current timetable has allowed a period of 2 months.

3. MATCHING THE CCS AND CENSUS REHEARSAL DATA

3.1 Background

The 2001 matching exercise will involve a combination of automated and clerical matching. There are four key stages:

- Automatically link households and individuals where key details match exactly.
- Automatically link households and individuals where key details are very similar. Similarity is determined by probability weights. The higher the probability weight, the closer the agreement between two records.
- Automatically select similar pairs of records for a clerical decision on their matching status.
- Clerically search for Census records corresponding to any unmatched CCS records.

The matching process is described in detail in the paper ONS(ONC(SC))98/14.

3.2 Changes in Methodology

A small number of changes to the methodology outlined in the above paper have resulted from the ongoing research. These changes are as follows:

- The matching weights derived from the 1999 Census Rehearsal will be used as starting weights for the probability matching in 2001. These weights will be tuned to the data being matched.
- The algorithm for deriving the household structure has now been developed and will be tested as part of the 1999 Rehearsal Evaluation.
- Tenure will not be used as a matching variable since it is a key household analysis variable. Type of accommodation may be used instead. This evaluation will determine the precise matching variables to be used.

3.3 Differences in Methodology Specific to the Rehearsal

Prior to the analysis described here, it will be necessary to match the rehearsal data clerically. The weights calculated from this clerically matched data would be used as starting weights for the 2001 matching process. The rehearsal weights can be iteratively tuned to the data being matched. This pre-automated matching clerical matching process will not be repeated in 2001 and therefore is not included in this evaluation. For the purposes of the following evaluation, the clerically matched records are deemed to be 'true' matches.

3.4 Evaluation Topics

The evaluation of the matching process will focus on nine areas:

- Matching feasibility;
- Derived variables;
- Matching variables;
- Input data;
- Matching techniques;
- Matching software;
- Clerical matching training;
- Calculation / updating of matching weights;
- Timing

3.4.1 Matching Feasibility

Objective: To determine whether automated matching can be performed to requirements in 2001.

Outcome: Evaluation of the clerically matched data indicates that the matching can be performed within the necessary time and accuracy requirements.

For further details see ONS(ONC(SC))00/14.

3.4.2 Derived Variables

Objective: To evaluate the derived variables used in the Rehearsal matching and define the most appropriate derived variables for 2001 matching purposes.

Outcome: Evaluation ongoing.

3.4.3 Matching Variables

Objective: To select appropriate matching variables for the 2001 automated matching process.

Outcome: Evaluation ongoing.

3.4.4 Input Data

Objective: To determine the set of data that facilitates the most accurate matching.

Census and CCS data will be available in two forms:

- Load data
- Post DEIS data

Outcome: Practical considerations mean that the matching will be performed using load data. There has not been time to investigate the quality implications of this, but there is no reason to assume that they would be significant. Timing is felt to be the critical factor here.

3.4.5 Matching Techniques

Objective: To determine the most appropriate matching strategy for 2001.

Once the data have been analysed and practical experience gained of matching census data, it is sensible to review the matching strategy to determine whether it is appropriate.

Outcome: Evaluation ongoing.

3.4.6 Matching Software

Objective: To specify an automated matching system appropriate for linking the 2001 Census and CCS data.

Outcome: The Rehearsal matching system was very successful and received extremely positive feedback from users. It is proposed to use this prototype software as the basis for the 2001 software. Lessons learned from the Rehearsal need to be incorporated. This work is ongoing.

3.4.7 Clerical Matching Training

Objective: To produce comprehensive training materials for 2001 clerical matching.

Outcome: Evaluation of the training given for the Rehearsal and feedback from matchers is ongoing.

3.4.8 Calculating / Updating Matching Weights

Objective: To produce starting weights for the 2001 automated matching process and determine appropriate methodology for updating them as the matching process progresses.

The current matching proposal involves using weights calculated from the 1999 Rehearsal data as starting weights for the 2001 automated matching process. These weights will be updated as matched 2001 data becomes available.

Outcome: Evaluation ongoing.

3.4.9 Timing

Objective: To determine the time and number of people required to perform the matching in 2001.

Outcome: The current proposal is to recruit ten members of staff dedicated to matching for 2001. It is estimated that it will take approximately six days on average to match each Estimation Area with this level of staffing. This fits comfortably within the current timetable.

4. ESTIMATION

4.1 Background

The CCS is designed to produce direct estimates for age-sex groups at the design group level. First, a Dual System Estimation (DSE) method is used to estimate the number of people in different age-sex groups missed by both the Census and CCS within each postcode in the CCS sample. Second, the postcode level population counts obtained from these DSEs are used with a modified ratio estimator to obtain

final counts for the design group as a whole. This process is carried out separately in each Hard to Count category. The methodology for estimating the design group populations is given in more detail in ONS(ONC(SC))00/03A.

Many LADs will not have enough CCS postcodes to allow accurate direct population estimates to be made. Therefore, it is proposed to use synthetic (or small area) estimation to produce accurate LAD level population estimates. Synthetic estimation uses information from the whole design group to apportion the estimated undercount to the LADs. A LAD adjusted synthetic estimator will be used to derive the LAD estimates as detailed in ONS(ONC(SC))00/03B.

4.2 Evaluation Topics

The evaluation of the estimation processes will focus on the following four areas:

- The design group estimation software;
- The timing of the design group estimation;
- The LAD estimation software;
- The timing of the LAD estimation.

4.3 Status of evaluation

Due to delays in the delivery of the 1999 Rehearsal data and further development of the estimation methodology, the estimation using the Rehearsal data has not yet taken place. It is planned to carry out the estimation and evaluate the revised systems by the end of July 2000.

However, a dry run of the ONC processing systems was undertaken in November 1999. This contained about ½ million records and was used to assess the likely time required to perform the estimation using the prototype systems, and the outcomes are reported below.

4.3.1 Estimation Software

Objective: To evaluate the prototype design group estimation system and determine an appropriate software package for the development of the 2001 system.

Outcome: The dry run also established that SAS is an appropriate tool for development of the 2001 system, since it offers further functionality that is likely to be used for the final estimation systems.

4.3.2 Timing of Estimation

Objective: To determine the time required to perform the design group estimation in 2001.

Outcome: The entire process took approximately 30 minutes to run. It must be noted that this used a simple regression model for Design Group estimation and a synthetic estimator for producing LAD totals. However, it is expected that any impact from methodological changes will be cancelled by the use of higher specification hardware. Note that some additional time will be required

to check the estimation, and it is estimated that this may take an additional 90 minutes.

5. DEMOGRAPHIC ESTIMATES AND ADMINISTRATIVE RECORDS

5.1 Background

The 2001 Census-based population estimates will be compared to demographic estimates produced by Population & Vital Statistics (P&VS) Division. The strategy for the Quality Assurance of ONC estimates is given in ONS(ONC(SC))00/04; proposals and work to date on taking forward this strategy are reported in ONS(ONC(SC))00/18.

5.2 Evaluation Topics

The evaluation will focus on:

- Testing interfaces between processes, in particular the mechanisms for data flow;
- Quality of data from administrative records.

5.3 Status of evaluation

Due to delays in the delivery of 1999 Rehearsal data, it has not been possible to use Rehearsal-based estimates to develop and evaluate quality assurance processes. Instead, work is ongoing to develop quality assurance processes for a single Design Group (Southwest Hampshire) by comparing simulated census-based estimates against both demographic estimates and aggregate data from administrative sources. This work, however, is still focussed on the topics outlined above, and the outcomes to date are summarised below (for more detail, see ONS(ONC(SC))00/18).

5.3.1 Design Group Estimates

Objective: To evaluate the processes of determining whether the Design Group estimate is within a plausible range.

Outcome: It is proposed that plausibility ranges at sub-national levels (both Design Group and Unitary/Local Authority) are determined by (i) assuming that for similar types of areas, errors will be broadly constant between 1991 and 2001, and (ii) by using the differences between the demographic estimate and the other administratively-based comparators for a particular age-sex group as an indication of error.

5.3.2 Quality of Data from Administrative Records

Objective: To assess the quality and reliability of data from administrative records to evaluate whether they are of sufficient quality for making comparisons with the ONC estimates.

Outcome: There will be a set of comparators for each of the age groups within each Design Group/Local Authority. As well as the demographic estimates produced by P&VS Division, this portfolio of comparators will include aggregated data derived from administrative records. A range of possible

administrative sources have been evaluated, and the following will be included in the development of prototype quality assurance procedures:

- Birth Registration data (babies aged under 1);
- DSS Child Benefit data (children aged 0-15);
- DSS Retirement Pension data (adults aged 65 and over);
- FHSA data (all age groups);
- Schools Census data (children aged 5-14);
- HESA data (students in higher education); and
- DASA data (armed forces).

6. IMPUTATION

6.1 Background

The final stage of the ONC process adjusts the census database at micro-level for underenumeration in three main steps:

- Matched Census and CCS data is used to model the probability of being counted in the Census in terms of characteristics of individuals and households. The models are calibrated to the agreed LAD estimates and applied to all individuals and households counted by the Census in order to calculate their 'census coverage' probabilities.
- A donor imputation system uses the coverage weights to create records for non-responding households and individuals who were missed from counted households.
- Finally, the imputation of individuals and households is controlled to agreed LAD totals for household size and age-sex structure.

To date the imputation methodology has been evaluated using simulation studies. The imputation process is described in detail in the paper ONS(ONC(SC)99/08).

6.2 Changes in Methodology

There have been no changes to the core imputation methodology outlined in the above paper. Possible differences in the methodology for placing imputed dummy households are described in paper ONS(ONC(SC))00/17.

6.3 Differences in Methodology Specific to the Evaluation

The voluntary nature of the 1999 Census Rehearsal has led to a significantly lower response rate to that expected for the 2001 Census. Additionally there have been significant delays in the availability of the Rehearsal data. As a consequence a number of simulated censuses have been constructed to provide datasets for use in evaluation of the system, as described in paper ONS(ONC(SC))00/17.

6.4 Evaluation Topics

The evaluation of the imputation process will focus on the following six areas:

- Household and individual coverage weights;
- Household imputation;
- Individual imputation;
- Constraint of marginal totals;
- Imputation software;
- Timing.

An important component of the evaluation will be to ensure that the imputation system interfaces with other systems in a live environment.

6.5 Status of Evaluation

6.5.1 Household and Individual Coverage Weights

Objective: To assess whether household and individual coverage weights created are appropriate.

Outcome: Simulation and evaluation work has been delayed as a result of prioritising development of code to implement the use of dummy forms in household imputation. Preliminary results suggest that the system is capable of generating appropriate weights at the household level. Individual statistics have yet to be investigated.

6.5.2 Selection of Donor Households

Objective: To evaluate whether the method selects appropriate donor households.

Outcome: The donor households examined for the preliminary studies are able to closely reflect the estimated totals. Further work is ongoing.

6.5.3 Imputation of Households

Objective: To evaluate whether the system successfully adds those households selected for imputation into the database.

Outcome: Imputed households are correctly placed into the database. Work is ongoing to determine the level of improvement possible by using dummy forms to control the placement of imputed households (see paper ONS(ONC(SC))00/17)

6.5.4 Selection of Donor Individuals

Objective: To evaluate whether the system selects appropriate donor individuals.

Outcome: Evaluation of individual level imputation has been put back following the decision to prioritise the use of dummy forms in household imputation. The simulations and evaluation will now take place in late June and early July.

6.5.5 Imputation of Individuals

Objective: To evaluate the suitability of recipient households

Outcome: See 6.5.4 Selection of Donor Individuals.

6.5.6 Constrain Marginal Totals

Objective: To evaluate whether pruning and grafting process constrains margins.

Outcome: Evaluation of the pruning and grafting procedure to constrain marginal totals will be conducted as part of the evaluation of the individual level imputation in June and July.

6.5.7 Imputation Software

Objective: To identify what changes are required to the current prototype to develop a fully automated imputation system.

Outcome: The developments to the program are described in paper ONS(ONC(SC))00/17. The overall performance of the program is currently being evaluated, and will be further tested as part of a complete Rehearsal of the downstream processes.

6.5.8 Timing

Objective: To determine the timing and resources required to perform the imputation in 2001.

Outcome: Preliminary simulations suggest that imputing a single LAD Design Group of around 500,000 people will take between 12 and 24 hours. Design Groups containing more than one LAD may take slightly longer, but are likely to take less than 24 hours to complete.

7. ONC OVERVIEW

7.1 Background

This section considers aspects of the One Number Census that do not fit within the previous project areas. One important part of the evaluation involves ‘taking a step back’ from the individual processes and understanding the effect that the ONC as a whole has on the Census data.

Another key issue is that of Communal Establishments. The methodology for estimating the underenumeration within Communal Establishments is an area that has yet to be addressed in detail. This methodology will impact on all ONC processes, as well as other Census areas.

Interfaces with other Census project areas are also covered in this section. The ONC is highly dependent on work undertaken by other areas such as Field, CCS, IS, Processing, Data Quality and Outputs and it is therefore essential that proper communication channels be established. It is also important that GROS and NISRA are kept fully informed of ONC work to ensure a successful UK One Number Census.

Some data quality issues of particular relevance to the ONC are evaluated in this section.

Finally, ONC timetables and staffing plans for 2001 will need to be revised as a result of the evaluation findings. Any revisions will need to fit in with the overall Census timetable and budget.

7.2 Evaluation Topics

The overview evaluation will focus on the following five areas:

- Measurement of change of data through ONC processes;
- Communal Establishments;
- Project interfaces;
- Data quality;
- Timetable and staffing.

7.2.1 Measurement of Change of Data through ONC Processes

Objective: To measure and understand the changes in the dataset as it passes through the ONC processes and back to the Census database.

Outcome: Ongoing - The Rehearsal data have not been through the estimation and imputation processes so this has not yet been possible.

7.2.2 Communal Establishments

Objective: To determine an appropriate method for estimating the underenumeration within Communal Establishments in 2001.

Outcome: The Steering Committee meeting in February directed that Communal Establishments should not be covered. However small CEs included within the CCS postcodes will have to be included, since it is impossible to exclude them from the sampling frame and there is a grey area between the definitions of a household and a communal. It follows from this that we should assume no underenumeration or (more likely) the same level of underenumeration for CEs as for the general population. This is being taken forward in the implementation of the estimation methodology.

7.2.3 Project Interfaces

Objective: To ensure that interfaces are in place with other areas of Census to allow co-operation between different areas and organisations.

Outcome: Ongoing. The rehearsal of the CCS identified some interfaces that needed strengthening. These have been developed with an ONC presence built into the CCS project plans and weekly interface meetings held with ONC, GROS, NISRA and CCS teams.

Ensuring delivery of a database suitable for the evaluation of the ONC has been extremely useful in focussing communications with Census processing.

The draft specification for the processing of the CCS data in 2001 has been prepared in conjunction with Census specifications and sent to the contractor.

7.2.4 Data Quality

Objective: To assess the quality of the question responses and dummy form information collected in the Census and CCS rehearsals.

Outcome: The ONC matching process found that the capture process was creating 'phantom' people. Jennet Woolford has examined this issue further, it has been fed back to the contractor and proposals for dealing with the problem have been made.

A sample of CCS forms have been re-keyed and compared with the data captured by the contractor. Initial results indicate that the quality of the captured data at an item level is high.

Results from this exercise along with an examination of the quality of the Census data will be fed back to the contractor and further tests of the capture process will explicitly address the problems identified.

Work is ongoing to look at the differences in responses to questions between the matched Census and CCS data and to examine the quality of the dummy form information.

7.2.5 Timetable and Staffing

Objective: To ensure that any timetables and staffing plans for the ONC in 2001 are achievable, and to revise these in line with evaluation findings as necessary.

Outcome: Ongoing – however evaluation carried out so far indicates that the current ONC timetable is achievable and estimates of staffing requirements good. The timetable for ONC is being considered within the timetable for all census processes and a rehearsal of all systems is planned for end 2000.

Staffing profiles for 2001 are to be revisited and a strategy for recruitment developed.

REFERENCES

ONS(ONC(SC))98/12 ‘*Census Coverage Survey: The Precision of Population Estimates for Different Sample Sizes and Design Areas*’; Abbott, O., Brown, J., Chambers, R., Diamond, I., Dixie, J.

ONS(ONC(SC))98/14 ‘*One Number Census Matching*’; Baxter, J.

ONS(ONC(SC))99/08 ‘*A Donor Imputation System to Create a Census Database Fully Adjusted for Underenumeration*’; Steele, F., Brown, J. and Chambers, R.

ONS(ONC(SC))00/01 ‘*One Number Census Methodology*’; Diamond, I.

ONS(ONC(SC))00/03A ‘*Estimation Strategy for Design Group Estimates by Age and Sex for the Census Coverage Survey*’; Brown, J., Chambers, R., Cruddas, M.

ONS(ONC(SC))00/03B ‘*One Number Census Local Authority Estimation*’; Abbott, O., Brown, J., Chambers, R., Cruddas, M.

ONS(ONC(SC))00/04 ‘*A Quality Assurance and Contingency Strategy for the One Number Census*’;

ONS(ONC(SC))00/06 ‘*1999 Census Rehearsal: Evaluation Plan*’

ONS(ONC(SC))00/10 ‘*Design Groups for 2001*’; Wright, E.

ONS(ONC(SC))00/14 ‘*ONC Matching*’; Woolford, J.

ONS(ONC(SC))00/15 – ‘*2001 Hard to Count Index*’; Abbott, O.

ONS(ONC(SC))00/17 – ‘*ONC Imputation: update*’; Howell, D.

ONS(ONC(SC))00/18 – ‘*Quality Assurance: update*’; Wright, E, Diamond, I .