

ONE NUMBER CENSUS STEERING COMMITTEE**One Number Census Imputation - Progress Report and Simulation Evaluation Exercise.**

1. The One Number Census imputation system has had significant modifications since it was last evaluated. This paper outlines these modifications and describes the simulations being conducted to evaluate the performance of the ONC imputation system.
 2. Simulations have commenced to evaluate the ability of the system to recreate a known target population, and to investigate the possible benefits arising through the use of dummy forms in the imputation.
 3. Preliminary results suggest that using dummy forms to place imputed households results in an improvement in the agreement between the adjusted database and the true population.
- 3. Members of the Steering Committee are asked to note the paper**

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

June 2000

1. BACKGROUND

- 1.1 Imputation is the final stage of the One Number Census (ONC) which aims to adjust the 2001 Census database at micro-level for the estimated underenumeration. Briefly, this will be achieved by the following method:
- using matched Census and CCS data to model the probability of being counted in the Census in terms of characteristics of individuals and households;
 - the probability of being counted for each type of household and individual is then converted to a 'census coverage' weight;
 - the census coverage weights for individuals and households are calibrated to agree with the estimated LAD totals;
 - the coverage weights are used by a donor imputation system to create records to compensate for non-responding households and for individuals missed from counted households;
 - the totals are then constrained to the estimated LAD totals for household size and age-sex structure.
- 1.2 The processing block for the ONC is the Design Group. The Imputation System processes one Design Group at a time and passes a listing of the changes to be made to the Census database back to Census IS to implement. The Imputation System does not interface directly to the Census Database.
- 1.3 To date the ONC imputation methodology has been evaluated using simulation studies based on data from the 1991 Census, as described in paper ONS(ONC(SC)99/08). Ownership of the ONC imputation system passed from Southampton University to ONS in June 1999. Since this time there have been a number of changes to the imputation system in order to improve the speed and robustness of the system, and to convert from a prototype to an operational system. These changes are described in Section 2.
- 1.4 In order to ensure that the imputation system correctly adjusts for underenumeration it is important to re-evaluate the system following these modifications. The planned programme of simulations is presented in Section 3.
- 1.5 Information about the location of non-responding households is collected in 'dummy' forms (these are census forms created during enumeration for identified non-responding households – they cover absent households, refusals and vacant properties). Using this information to control the placement of imputed households has the potential to reduce the discrepancy between the final adjusted Census database and the 'true' population. Simulations to evaluate the effect of using dummy form information in the imputation are described in Section 4.

2. CHANGES TO IMPUTATION SYSTEM SINCE PAPER ONS(ONC(SC)99/08)

- 2.1 There have been a number of changes to the ONC imputation system since the previous evaluations were conducted. These changes have been made to improve the speed of the system, to eliminate minor bugs, and to improve the robustness of the system. Changes have also been made to convert the models used to the required format for the 2001 Census.
- 2.2 The majority of these changes have been to improve the speed of the ONC imputation system. The original system required approximately 48 hours to process a test dataset. Rewriting sections of code has resulted in a several-fold decrease in run time, despite extra features having been introduced. During the process of rewriting the system a number of minor bugs were detected and corrected. In addition to changes to the code for efficiency reasons, changes have been made to adjust the models within the ONC imputation system to use 2001 data rather than the 1991 data the system was developed on. The model has also been adapted to use 5 year age bands across all ages.
- 2.3 The final stage of the ONC imputation procedure is to constrain the final totals so that the household size and age/sex totals exactly meet the estimated values. In order to meet household size requirements it is necessary to add or remove imputed individuals to move some households between size categories. While doing this the system ensures that imputed people are deleted in such a way that the final dataset exactly hits the age/sex targets from the LAD estimation.
- 2.4 The original prototype system could not be guaranteed to produce a final dataset consistent with the estimated totals. Furthermore, this failure could only be identified by observing that the program continued to run indefinitely. This section of the system has therefore been totally redesigned. The chance of failure has been reduced, with no failures noted in testing to date. As an additional backup the section capable of failing has been made as small as possible, thus making it easy to re-run any failures in a matter of minutes rather than hours. Any failure is down to random chance, and thus re-running the relevant section of code is likely to resolve any problems. The program will conduct such re-runs automatically.
- 2.5 Full details of the system are given in the ONC Imputation System Manual, which is available on request.

3. SIMULATION METHODOLOGY

- 3.1 The evaluation of the ONC Imputation Methodology and Computer System has two main steps:
1. the statistical evaluation which concerns comparing the post-imputation database against the known 'truth'; and
 2. the operational evaluation which concerns the practical issues surrounding the implementation.
- 3.2 The 1999 Census Rehearsal data is not suitable as the sole dataset for testing the system. Firstly, the 'true' population is not known, and thus the accuracy of the imputation cannot be assessed. Secondly, the response rates to both the Census and CCS were significantly lower than is expected for the 2001 Census. Finally, delays in processing rehearsal data have made it imperative that testing commence before the data was available.
- 3.3 Artificially constructed simulations are therefore being used to undertake a statistical evaluation of the system, and are described in this paper. An operational evaluation will also be undertaken as part of the Census Rehearsal to further assess the practicalities of the implementation of the ONC imputation system.
- 3.4 The data used to conduct the evaluation comprises an extract of post-edit 1991 Census data, selected with the intention of forming a reasonably representative sample of the UK population. These comprise ten simulations from each of two different area types:
- A. Mostly Urban
 - B. Mostly Rural
- 3.5 The simulated Censuses used in this test are based on 1991 census returns, and have been recoded to 2001 categories in order to provide relevant test of the system. In each case the actual Census database has been treated as the 'true' population, and simulated census and CCS datasets have been constructed as subsets of this true data. Probabilities have been assigned to each household and each individual to represent their likelihood of being missed in any census and CCS. Further checks have been conducted to ensure that resulting simulated households have at least one adult present.
- 3.6 Ten simulated censuses have been constructed for each of the two design groups. For each simulation the ONC imputation system is run, using the 'true' values from the recoded 1991 census as the 'target' estimates to impute to. The process thus provides a test of the system's ability to impute to known targets, and it is possible to ascertain how close the imputation system came to 'reconstructing' the actual population.
- 3.7 There are a number of different factors to be analysed in running the simulations:

- Ability of the imputation to correctly impute to the age/sex and household size estimate at the Design Group/LAD level
 - Ability to achieve a realistic distribution of imputed individuals and households across the Design Group.
 - Ability to achieve a reasonable distribution of other (non-constrained) variables
 - Robustness of the system
 - Speed of the system
- 3.8 The suite of 20 simulations described above will be run twice. Initially the system will treat each design group as containing a single LAD. This is comparable with the simulations conducted by the University of Southampton on the prototype system. A second set of simulations will then be carried out on the same datasets, where imputation will be conducted respecting the true distribution of the LADs within the Design Groups. The urban and rural Design Groups contain 3 and 5 Local Authorities respectively.
- 3.9 This dual approach has two benefits. Firstly, by running simulations comparable with the previous studies the results can be compared. Secondly, the adjustments required to run the system on multiple LADs can be evaluated separately from the adjustments already made to the system.

4. POSSIBLE USE OF DUMMY FORMS IN ONC IMPUTATION

- 4.1 The number and characteristics of missed households can be estimated at the Design Group and/or LAD level using CCS and Census data. These missed households must then be allocated to physical localities within that Design Group. The distribution of missed households within a Design Group can be estimated by using 'hard to count' values for different postcodes, but this *a priori* method cannot account for unexpected clusters of missed households. Such clusters may occur as a result of localised difficulties during the Census.
- 4.2 Dummy forms provide information concerning the possible location of missed households within a Design Group, and therefore it is anticipated that dummy forms may identify some of the areas in which such unexpectedly high clusters of undercount occur. Consequently the use of dummy forms to control the physical location of imputed households has the potential to improve the accuracy of the distribution of the final adjusted dataset.
- 4.3 Two alternative methodologies exist for placing imputed households into the adjusted database, either using or not using dummy forms. If dummy forms are not used then households are imputed into a random postcode within the ED of the duplicated donor household. This ensures that all households are located in a suitable area, but does not control the placement of imputed households at the postcode level. All imputed households are given a physical location identical to the centroid of their recipient postcode.

- 4.4 Dummy forms can be used to control the precise placement of imputed households, and thus improve the distribution of households at a local level. The system uses a three stage imputation system. In the first stage households are placed in available dummy forms in the same postcode as the donor household. Not all households will find a suitable recipient household by this process, so a second stage takes all unplaced households and attempts to find a dummy form somewhere in the Design Group which has a similar household (same tenure, hard to count, agestructure and ethnicity) in the same postcode. If more than one suitable recipient form is located then the geographically closest is used. Finally any households remaining unplaced after this second stage are allocated to a random postcode within the same ED as the donor household.
- 4.5 The benefits to be gained by using dummy forms are currently being evaluated by repeating several of the Design Group level simulations in which the missed households have been converted into dummy forms. Because the actual distribution of the original population (the recoded 1991 census data) is known, it is possible to assess the level of improvement possible if 100% accurate dummy forms are available for use in the imputation procedure.
- 4.6 Preliminary results from the simulation exercise are available to ascertain the impact of using dummy forms to control household placement in the imputation process. The difference between the adjusted database and the true 'target' population are computed for each Enumeration District. The standard deviation of these differences across the entire Design Group has then been computed to indicate the level of error between the adjusted and target populations. This analysis is repeated at the postcode level. The standard deviations for the hard to count index, tenure, and the number of cars per household are presented in Tables 1, 2 and 3 respectively.
- 4.7 The distribution of the hard to count score is improved the most, this is unsurprising since hard to count is distributed spatially at the ED level. Other variables which show some degree of spatial variation, such as Tenure, are also improved, although not to the same extent that the hard to count distribution is. The distribution of the number of cars per household is also improved by the use of dummy forms, even though this variable is not directly considered within the model.
- 4.8 Using accurate dummy forms to control the household imputation clearly improves the distribution of households within a Design Group, producing a final adjusted dataset which is closer to the original 'true' population than if dummy form information is not used. Further simulations are ongoing to produce a larger sample size for the results than the single test presented here, and to investigate individual characteristics as well as household level ones.

Table 1: Hard To Count Index

	At the ED level		At the PC level	
	With Dummy Forms	Without Dummy Forms	With Dummy Forms	Without Dummy Forms
Mean number of households of hard to count category per area (ED or PC)	162.56		15.56	
Standard Deviation of differences between target and adjusted populations	1.15	4.98	0.36	1.16

Table 2: Tenure Category

	At the ED level		At the PC level	
	With Dummy Forms	Without Dummy Forms	With Dummy Forms	Without Dummy Forms
Mean number of households of each tenure category per area (ED or PC)	44.74		6.52	
Standard Deviation of differences between target and adjusted populations	1.66	2.26	0.54	0.73

Table 3: Number of cars per household

	At the ED level		At the PC level	
	With Dummy Forms	Without Dummy Forms	With Dummy Forms	Without Dummy Forms
Mean number of households of each number of cars category per area (ED or PC)	46.84		5.59	
Standard Deviation of differences between target and adjusted populations	1.51	2.20	0.50	0.66