

ONE NUMBER CENSUS STEERING COMMITTEE

Changes to the ONC Imputation System

1. This paper outlines the modifications that have been made to the implementation of the ONC imputation methodology in light of the 2001 Census and CCS data available. It was found that when implemented directly at the Local Authority District level, the methodology did not produce stable results.
2. The implementation of the methodology has therefore been revised to provide stability whilst retaining the statistical integrity and quality of the database.

The Steering Committee are asked to note the paper.

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

February 2002

Changes to the ONC Imputation System

Owen Abbott and James Brown

1. Introduction

- 1.1 This paper outlines the modifications that have been made to the One Number Census (ONC) imputation system in light of the 2001 Census and CCS data available. A full description and cause of the problems encountered are given, as well as the thinking behind the solutions.
- 1.2 The solution evolved gradually as at each step it was necessary to understand and analyse the impact the changes made before implementing further changes to the system.

2. Imputation Methodology Summary

- 2.1 The ONC imputation methodology can be split into three phases:

Phase 1: Missed Household Imputation

- 2.2 The first phase is the modelling and imputation of households estimated to have been missed by the Census. This uses the matched Census and CCS dataset to derive coverage weights for households, which are then calibrated to ONC Tenure estimates at the Local Authority District (LAD) level. The weights are used to then determine which households to impute, and the households are placed into either Census dummy forms or into a random postcode within the LAD.

Phase 2: Missed persons within counted household imputation

- 2.3 The second phase is the modelling and imputation of persons within counted households estimated to have been missed by the Census. Again, the matched Census and CCS data are used to derive coverage weights for persons missed within counted households, and these weights are calibrated to the age-sex LAD ONC estimates. The imputation of these persons is tightly controlled and the imputations are generally of a high quality.

Phase 3: Pruning and Grafting

- 2.4 The third phase is known as Pruning and Grafting, which adjusts the imputed database to ensure that household size targets and Local Authority District (LAD) by age-sex group ONC estimates are met exactly. However, this last process reduces the quality of the imputed data as it is not as controlled as the previous stages - it is only constrained by the household size and age-sex ONC LAD estimates. Furthermore, the simulation work indicated that there is a risk that a solution cannot be found.
- 2.5 Further details of the methodology can be found in ONC Steering Committee paper ONC(SC)99/08.
- 2.6 This methodology was, at the research stage, implemented at Design Group level (approx 500,000 persons). This prototype achieved a fully imputed database that met the Design Group population estimates using a simulated set of Census and CCS data.
- 2.7 The Local Authority District estimation methodology presented in Steering Committee paper ONC(SC)00/03B was agreed in February 2000. With regard to imputation it was subsequently

decided within the ONC project that the 2001 Imputation system would be implemented to carry out the imputation within each LAD separately, although the models would be fitted at the Design Group level.

3. Issues in adopting methodology

3.1 The first Design Group that was made available for live processing contained four LADs, each of a reasonable size. The undercount patterns that were estimated were not particularly different from those experienced in the 1991 Census. Therefore in terms of undercount the area was as expected and there was no reason to suspect that the Imputation methodology would not work when implemented at the LAD level.

3.2 However, after completing the modelling and imputation of households estimated to be missed in the Census for each LAD, a comparison was made of the age-sex counts of the imputed persons within imputed households against the gold standard ONC estimates of how many extra people were required. In a number of cases, particularly for one LAD the household imputation had put in far too many people (up to one hundred extra persons in some age-sex groups within an LAD when only twenty or so were required). Three overriding reasons for this over-imputation were discovered. These are discussed below:

a) Household Estimates at LAD level

3.3 The household undercount pattern by LAD was very different from the person undercount pattern by LAD. The household coverage for one particular LAD was estimated to be 91% whereas the person estimates indicated coverage of 97%. As a result, a lot of households were imputed and far too many people were added. The pruning and grafting phase is designed to remove any of these excesses, but this over-imputation was too large and there is a restriction that all people within an imputed household cannot be removed (as the household would then be empty).

3.4 Household undercount is determined by ONC Estimation purely for use in the imputation process as a set of control totals. Further analysis determined that while the methodology used to derive the LAD level estimates was working well for age-sex groups, the LAD household estimates did not appear to be sensible. This may be because the age-sex groups are the level for which the methodology was designed and researched. The household level tenure estimates are more difficult to quality assure at LAD level as no reliable comparative data exist.

b) Consistency of coverage weights

3.5 The household undercount model includes a variable that explains the structure of the persons within the household (e.g. young couple, family etc) and thus includes an implicit age-sex undercount. For age-sex groups where very little undercount was estimated, there was no restriction on the household weights to ensure consistency.

3.6 This problem was most evident in the older age-sex groups where the population is smaller (and thus the sample size is small) and the undercount is small in numerical terms. In particular, there is no constraint on the consistency of the weights for households containing one or two elderly people to ensure they are consistent with the age-sex undercount estimates for the elderly. As a result, too many households containing elderly persons were imputed.

3.7 Whilst an investigation of the consistency of the household weights with age-sex totals was desirable during the ONC research, due to resource constraints and other areas of imputation research taking priority it was not undertaken. The prototype imputation system used much wider age-sex bands (in particular a 45-79 year old grouping) and therefore this type of problem would not have been apparent. However, the 2001 system uses five year age bands throughout and thus at this finer level of imputation there are more problems with consistency.

c) No restriction on persons imputed in households

3.8 For age-sex groups where very little undercount was estimated, there was no restriction on the household imputation to avoid imputing households with these types of persons. This was most evident in the older age-sex groups where the population is smaller and the undercount is small in numerical terms. However, the problem could potentially occur across any of the age-sex groups.

3.9 The impact of not having any restriction is that we expect pruning and grafting to sort out the age-sex distribution by removing these over-imputed persons. However, this later process reduces the quality of the imputed data and there is always the risk that a solution cannot be found. It is therefore preferable and sensible to get it right in the first place when there is more control over the characteristics of the households imputed.

4. Solutions

4.1 Household Estimates at LAD level

4.1.1 The first problem is the derivation of LAD household estimates by the ONC Estimation. Because these estimates were determined to be unreliable, it was decided not to use LAD level household estimates as targets and to only use Design Group level targets. This is the level at which the prototype system was researched.

4.1.2 This change would firstly allow the imputation weights derived from the model to determine how many households it should impute into each LAD rather than providing a hard target. It would secondly improve the overall quality of imputation at Design Group level, as we would no longer be ensuring we meet LAD household size estimates (which requires a lot of pruning and grafting). However, we would not be controlling the household size distribution at LAD level and therefore a check will have to be carried out after imputation to ensure the distribution is sensible.

4.1.3 Once implemented, this reduced the level of over-imputation within the LADs although not by much. It also reduced significantly the overall number of prunes and grafts that were required. This is a good indication that this change will not have a detrimental effect on the quality of the imputation, and may well improve the overall database quality.

4.2 Consistency of coverage weights

4.2.1 In order to ensure that there is consistency between certain coverage weights and the corresponding age-sex targets, a check on the household coverage weights was added. This determines if the weights for single and couple elderly households will result in an over-imputation of these age-sex groups. If it does, then the weights for this group are reduced to

ensure consistency. This is also implemented for young single and young couple households. Other checks are not possible due to the structure variable being quite broad.

4.2.2 This additional constraint reduces the risk of there being any instability in the weights - for instance if the weights have overestimated the coverage of a particular type of household. This is likely where there are small sample sizes (e.g. elderly type households will be particularly vulnerable). Again, this change did reduce the over-imputation slightly as the weights were more consistent, but not to the extent such that the amount of Pruning and Grafting was reasonable.

4.3 No restriction on persons imputed in households

a) Preventing imputation within particular age-sex groups

4.3.1 In order to prevent the imputation of persons of a particular age-sex group, households where all persons are of an age-sex group for which no undercount was detected are flagged and excluded from the household imputation - we effectively give these households a coverage weight of 1.

4.3.2 By preventing this type of household from being imputed, we further reduce the risk of not being able to meet the age-sex target. At present, we have to prune all over-imputed people out of the database, and the pruning will not converge if, for instance, they are all in single person households. This will therefore improve the imputation quality by reducing the amount of pruning of people in imputed households.

4.3.3 Once implemented, this reduced the over-imputation to zero for those age-sex groups with no undercount. However, it increased slightly over-imputation for other age-sex groups to compensate, as we still have to impute the same number of households.

b) Extending the donor household search based on age-sex totals

4.3.4 At present, the imputation methodology applies the coverage weights to each of the Census households, and then sorts them by ascending weight. Cumulative sums of both the weighted and unweighted counts are maintained. When the imputation weights get 0.5 higher than the unweighted census count, we look for a donor household to copy based on the location (e.g. LAD, Enumeration District) and characteristics (Tenure, Hard to Count, Ethnicity and Household Structure) of the household.

4.3.5 The most significant change to the methodology was to add a check to see whether the potential donor households have any persons of an age-sex group where we have already imputed enough persons. If all potential donors have such a person, then the household characteristics and location search criteria are gradually relaxed until a household is found that does not contain any such persons. Annex A describes how the criteria are relaxed.

4.3.6 Once such a household is found, the household is imputed and the age-sex counts of persons to be imputed are updated. Thus as we get closer to the age-sex targets, the household searches will generally be wider and the quality of imputation will be poorer - we will not be imputing the best possible match on household type and location.

4.3.7 To reduce the impact this will have on the imputed database, the ordering of the imputations has been altered. Instead of finding a donor immediately when the weights indicate a household should be imputed (this is done in order of ascending weights), the household

characteristics required are simply stored in a cumulative list. This results in a list of all the households we wish to impute. This list is then re-ordered so that the highest weight households will be imputed first where there is a higher probability of imputing the best possible household. The relaxing of the search criteria will then generally only occur to low weight imputations when we are nearer to the age-sex by LAD targets.

4.3.8 As a result of carrying out the imputation in this order, the high weight households (i.e. the households we missed the most of) will have a better imputation quality than the low weight households. This is driven by the philosophy is that it is more desirable to get the high weight household characteristics correct.

4.3.9 This change had the largest impact upon the imputation results. Once implemented, the over-imputation was eliminated, and the results are analysed in section 5. However, for each area there will have to be a careful analysis of the number of times the search criteria were relaxed in order to obtain an indication of quality.

4.4 Pruning and Grafting Convergence

4.4.1 Simulation work suggested some convergence problems with the prototype pruning and grafting system. As a result, early development work at ONS concentrated on the pruning and grafting methodology and this work has been ongoing. Non-convergence can occur in two situations.

a) Over imputation of single person households

4.4.2 Non-convergence can occur if we have over-imputed an age-sex group and those over-imputations are single person households. Under the current methodology, these persons cannot be removed as we would be left with empty households.

4.4.2 To prevent this from causing non-convergence, we propose to treat these cases as households that should never have been imputed. In these cases the whole household will be removed and a replacement imputed household selected where we need an extra person of a particular age-sex group. In other words we will prune and graft single person imputed households.

4.4.3 However, this will be the final process to be used if the usual pruning and grafting method does not converge. It is anticipated that this will not be used greatly, but will be a useful backup in certain situations.

b) Too much Pruning and Grafting

4.4.4 If the amount of pruning and grafting is so large that we run out of recipient households, then the system will not converge - particularly for large household sizes. This is mainly driven by the difference between the age-sex defined population total and that implied by the household size distribution. If the difference is large (generally the age-sex estimates are higher), the system will carry out a lot of pruning and will then maintain the population totals by imputing a large number into households of size 6 or greater.

4.4.5 To prevent this from happening, we propose that the household size estimates need not be met exactly. If we were to construct confidence intervals around these estimates, they are likely to be larger than 2% of the estimate. Therefore the system will attempt to minimise the amount of pruning and grafting subject to obtaining a database that is within 1% of the household size estimates.

5. Results

5.1 **Table 1** shows the outcomes for the first Design Group when the imputation methodology was applied at LAD level without the changes discussed in section 4. The areas shown have been anonymously labelled.

Table 1 - Summary of over-imputation for imputation system applied at LAD level.

Local Authority District	Households imputed	Number of age-sex groups over-imputed	Total persons over-imputed	Highest over-imputation within age-sex group	Number of age-sex groups over-imputed where no undercount
1	880	12	250	55	7
2	1989	6	368	140	4
3	2041	21	1237	132	6
4	1393	11	228	55	6
Design Group total	6303	50	2083	140	23

5.2 **Table 1** shows the levels of over-imputation within each of the LADs. LAD 3 has the highest number of over-imputed persons occurring within the most number of age-sex groups (the total number of age-sex groups is 37). This area had a relatively low undercount of persons but the household estimates showed a high undercount - resulting in too many persons being imputed within wholly imputed households. Although LAD 2 has some over-imputation, this is mainly within the 4 age-sex groups where no undercount was estimated.

5.3 **Table 2** shows the outcomes for the first Design Group when the revisions outlined in **Section 4** were implemented. The last column shows the number of households excluded from the imputation due to the change outlined in **Section 4.3a**. This shows that some of the LADs will have had greater problems with the original pruning and grafting system, as there are more households where all adults could not have been pruned.

Table 2 - Summary of over-imputation for revised imputation system.

LAD	HHs imputed	Number of age-sex groups over-imputed	Total persons over-imputed	Number of households where all adults are of age-sex groups with no undercount
1	1037	1	1	1752
2	2202	0	0	2
3	1344	1	1	584
4	1720	0	0	2
Design Group total	6303	2	2	2340

5.4 **Table 2** shows that for the changed version, there is virtually no over-imputation. The implemented changes have redistributed the imputed persons into age-sex groups where we have sufficient missing persons. The 2 persons who have been over-imputed are cases where

there are more than 1 person of an age-sex group in a household, and that household is imputed - thus the imputed totals then exceed the target by 1. This is acceptable, as these additional people can be pruned out at a later stage.

- 5.5 The change in the number of households imputed into each LAD can be derived from **Tables 1 and 2**. The largest change is in LAD 3, where we have imputed 697 fewer households. These have been distributed throughout the other 3 LADs, with the majority (47%) going into LAD 4. **Table 3** shows that this redistribution is not driven by the coverage weights. The households of donor type 8 and higher (the cells are greyed for clarity) are those households that could not be imputed within another LAD. Thus the 652 households of type 8 in LAD 2 will be cases where a suitable donor could not be found in another LAD, and so a household was imputed into LAD 2.

Table 3 - Number of donor types by Local Authority District for revised imputation system (See Annex A for definition of donor type).

Donor Type	LAD 1	LAD 2	LAD 3	LAD 4	Design Group total
0	759	1451	1067	1321	4598
1	126	59	100	122	407
2	126	26	124	111	387
3	2	3	2	0	7
4	6	1	2	0	9
5	4	1	4	2	11
6	2	0	1	3	6
7	12	1	2	46	61
8	0	652	42	115	809
9	0	0	0	0	0
10	0	8	0	0	8
Total	1037	2202	1344	1720	6303

- 5.6 **Table 3** also gives an indication of the reduction in quality of the imputations. Overall, 73% of imputed households were of the best quality, and only 13% of imputations were carried out across the whole Design Group. Of the searches within LADs (i.e. up to donor type 7), only 94 imputations required the hard to count and household ethnic constraints to be relaxed. The majority simply required a widening of the geographical element of the search for donor households.
- 5.7 This analysis indicates that there has not been a significant loss in quality of the imputations - 73% were to the same high standard, and of the remaining 27% the majority were cases where there was an extended search across the whole Design Group rather than within the particular LAD.

6. Conclusions

- 6.1 The problem of over-imputation when applying the agreed ONC Imputation methodology at Local Authority District levels has necessitated a revision to the way in which the methodology is applied. This paper has outlined the analysis carried out to fully understand the problems, and the resulting solutions.

6.2 The adjustments to the system have resulted in the elimination of the over-imputation problems with a minimal impact on the quality of the imputed households. The system retains enough flexibility to be able to cope with all circumstances and the changes have virtually eliminated the risk of the pruning and grafting failing to converge. Analysis will continue for each area as it is processed to ensure the quality of imputation is maintained at a high level.

ANNEX A

Definition of search criteria for selection of donor records when imputing households

Donor type	Criteria for selection of donors
0	Select from households in the same Enumeration District, Tenure, Household Structure, Hard to Count, Ethnicity.
1	Select from households in the same Ward, Tenure, Household Structure, Hard to Count, Ethnicity.
2	Select from households in the same LAD, Tenure, Household Structure, Hard to Count, Ethnicity.
3	Select from households in the same Enumeration District, Tenure, Household Structure, Hard to Count.
4	Select from households in the same Ward, Tenure, Household Structure, Hard to Count.
5	Select from households in the same LAD, Tenure, Household Structure, Hard to Count.
6	Select from households in the same Ward, Tenure, Household Structure.
7	Select from households in the same LAD, Tenure, Household Structure.
8	Select from households in the same Design Group, Tenure, Household Structure, Hard to Count, Ethnicity.
9	Select from households in the same Design Group, Tenure, Household Structure, Hard to Count.
10	Select from households in the same Design Group, Tenure, Household Structure.