



ONS(ONC(SC))98/10

ONE NUMBER CENSUS STEERING COMMITTEE

One Number Census: A consistent household/individual imputation strategy

1. This paper outlines some preliminary ideas for an alternative imputation process to that described in ONC(SC)98/07. The method is specifically designed to achieve consistency across the hierarchical structure of the data, ie households and individuals.
2. **The Steering Committee are asked to:**
 - a) **note the paper;**
 - b) **provide any comments at the meeting on the 27 April 1998, or in writing by 10 May 1998.**

**Tim Jones
Methods and Quality Division
Office for National Statistics**

**Room D2/10
1 Drummond Gate
London SW1V 2QQ**

April 1998

**One Number Census:
A consistent household/individual imputation strategy**

by Ray Chambers and Tim Jones

1. Introduction

The paper ONC(SC)98/07 by Marie Cruddas and Ray Chambers sets out an imputation-based approach to the final stage of the One Number Census process, based on multilevel modelling down to small areas for individuals. However, it is difficult with this approach to maintain consistency between estimates relating to household characteristics and estimates relating to individuals.

This paper outlines some preliminary ideas for an alternative imputation process specifically designed to achieve consistency across the hierarchical structure of the data, ie households and individuals. In order to do this, however, the non-coverage probabilities used in the process may be relatively crude.

The paper is in two main parts. In section 2 below, a general framework for achieving consistency is set out, beginning with some background and assumptions. In section 3, a procedure is developed for implementing such a consistent system for imputing individuals in a hierarchical framework. Finally there are some indications of the further work that will be required.

2 A consistent hierarchical framework

2.1 Background and assumptions

Although identifying the number and characteristics of individuals may be a top priority for the Census, their immediate social context (ie household characteristics) is also crucial (eg children living in households with a single parent; the number of persons living alone etc). Households are the fundamental units of enumeration for most social surveys. It is therefore assumed that consistency between the data at household level and that at individual level in the Census is a high priority.

In this note the term “household” is used throughout but this is not defined. It might as well be an address. It is however assumed that, below postcode level, it is the only level in the hierarchy above the individual. It may be further assumed (the hard part for the ONC) that it is the level at which relationships between the individuals are to be defined. (This paper does not cover individuals not living in households).

Suppose H_k denotes the number of households of size k in a given area. Then in a consistent system $\sum kH_k = N$, the population of the area.

In the work to date on weighting and imputation, the initial model assumes that the true population consists of those counted in the Census, together with those counted in the CCS that were missed in the Census. This assumption is unlikely to hold, and so an adjustment procedure is introduced to bring the results into line with control totals. The present proposals take a similar route. They may however affect the procedure envisaged in earlier stages in the process, and in particular the way in which the CCS/Census data are processed.

2.2 *The general framework*

It is convenient (as in previous work) to distinguish between households that have been missed completely in the Census, and those in which some individuals have been missed.

In the CCS areas, let m_k be the number of households of size k enumerated in the Census. For each k , there will be an extra h_k complete households identified in the CCS but missed by the Census, implying a household undercoverage probability of $h_k/(m_k + h_k)$ for households of size k . This is case (e) in the Cruddas/Chambers model. The strategy would be to identify a set of specific household characteristics, defined by the values of a categorical “household type” variable T say, such that similar-sized Census households are “exchangeable” with “extra” CCS households of the same type. In non-CCS areas, Census households of each type can then be randomly selected and used to impute for missing households of the same type in these areas using the type and size specific household undercoverage probabilities observed in the CCS areas. For $k=1$ the procedure would be at its simplest. In principle it is feasible to extend this to arbitrary values of k , but the number of possible categories for T may quickly become large.

In the second case, (c) in the Chambers/Cruddas paper, the CCS finds j extra people in F_{jk} households of size k . This means that in CCS areas there were F_{jk} too many households of size k found in the Census and F_{jk} too few of size $k+j$. In non-CCS areas it would be necessary to identify an equivalent number of households of size k , with similar characteristics to those F_{jk} cases in CCS areas, and to replace them with the same number of households of size $k+j$, in such a way that the extra people have similar characteristics to those found in the CCS cases. The imputation procedure would need to preserve the total number of households in any given area.

In order to select households into which extra people would be inserted, a sampling procedure would be used. The number of households to be selected in each small area would have to be determined, as would the probability of selection of any particular household (which would vary according to its characteristics). If the sampling were done sequentially (area by area) the probabilities could be varied dynamically to take account of previous selections in arriving at the appropriate control totals. Imputees would then come from a near neighbour in the CCS database.

3. A strategy for implementation

This is a two-step imputation procedure. In the first step we impute complete households in order to adjust for individuals in households completely missed by the Census. In the second step we impute individuals in households counted by the Census in order to adjust for individuals missed within these households.

To start, suppose no households are missed by the Census. The problem then is one of imputing missed individuals in households counted by the Census. As in Cruddas/Chambers, we assume that all imputation operations are restricted to households within a spatially defined “processing block”. However, in this case this block is quite large, e.g. a large Local Authority District, or a collection of contiguous smaller ones. Furthermore it is assumed that all households in such a block can be allocated to one of a finite number of household “types”, which we index by T . The definition of household type is not addressed here, but in general this categorisation will be based on an analysis of the CCS data which allows one to distinguish between households that are qualitatively different in their census coverage characteristics. Essentially the “type” of a household is equivalent to the “Impute Class” of an individual in Cruddas/Chambers. We also assume that

1. There is no overcoverage in the Census; and
2. A “CCS value” corresponds to the result of a matching operation which links data collected in the Census with data collected in the CCS operation, and a consequent reconciliation of these values. Thus, the “final” CCS value is always as large as the corresponding Census value.

Define F_{jkAT} as the number of households of type T in CCS locations within a processing block that were recorded in the Census as having k individuals in age-sex category A , but were discovered in the CCS as having $k + j$ ($j \geq 0$) individuals in age-sex category A . Denote the maximum value of j for individuals in age sex category A in CCS households of type T in the processing block as J_{AT} , and the corresponding maximum value of k as K_{AT} . That is, for CCS locations within the processing block, the Census found a maximum of K_{AT} individuals of age-sex category A in households of type T , and the CCS found a maximum of $K_{AT} + J_{AT}$ individuals of age-sex category A in households of type T .

To illustrate, suppose for particular values of A and T , $K_{AT} = 4$ and $J_{AT} = 2$. Dropping the A and T subscripts for the sake of clarity, we can cross-classify all households of type T in CCS locations in the processing block according to their Census and CCS counts of individuals in age-sex category A . This cross-classification will take the form:

Census Household Size	CCS Household Size						
	0	1	2	3	4	5	6
0	F ₀₀	F ₁₀	F ₂₀	0	0	0	0
1	0	F ₀₁	F ₁₁	F ₂₁	0	0	0
2	0	0	F ₀₂	F ₁₂	F ₂₂	0	0
3	0	0	0	F ₀₃	F ₁₃	F ₂₃	0
4	0	0	0	0	F ₀₄	F ₁₄	F ₂₄

Note that the total number of individuals in age-sex category A in households of type T found by the census in CCS locations in the processing block is then

$$n_{AT}(\text{Census}) = (F_{01} + F_{11} + F_{21}) + 2(F_{02} + F_{12} + F_{22}) + 3(F_{03} + F_{13} + F_{23}) + 4(F_{04} + F_{14} + F_{24})$$

while the corresponding count for the CCS is

$$n_{AT}(\text{CCS}) = (F_{10} + F_{01}) + 2(F_{20} + F_{11} + F_{02}) + 3(F_{21} + F_{12} + F_{03}) + 4(F_{22} + F_{13} + F_{04}) + 5(F_{23} + F_{14}) + 6F_{24}.$$

It is straightforward to see that $n_{AT}(\text{CCS})$ is always as large or larger than $n_{AT}(\text{Census})$. In general

$$n_{AT}(\text{CCS}) = \sum_{k=1}^{K_{AT}+J_{AT}} k \sum_{i+j=k} F_{jiAT}$$

while

$$n_{AT}(\text{Census}) = \sum_{k=1}^{K_{AT}} k \sum_{j=0}^{J_{AT}} F_{jkAT}.$$

A simple imputation procedure based on this framework proceeds as follows. For an arbitrary non-CCS location household of type T in the processing block that recorded k individuals in age-sex category A at the Census, define the probability of adding an extra $j \geq 0$ individuals in age-sex category A to this household as

$$\pi_{jkAT} = \frac{F_{jkAT}}{\sum_{i=0}^{J_{AT}} F_{ikAT}}, j = 0, \dots, J_{AT}.$$

Suppose now there are $M_{kAT}(\text{Census})$ type T households in the processing block that were “found” in the Census to have k individuals in age-sex category A. Of these, $m_{kAT}(\text{Census})$ were in the CCS locations in the processing block. By definition

$$m_{kAT}(\text{Census}) = \sum_{j=0}^{J_{AT}} F_{jkAT}$$

From the remaining $C_{kAT} = M_{kAT}(\text{Census}) - m_{kAT}(\text{Census})$ “non-CCS” households in this group, we now select households to which we “add” individuals. Operationally this can be accomplished by the simple expedient of independently allocating each of these C_{kAT} households a random number between zero and one, sorting them on the basis of these random numbers, and then, for $j = 1, \dots, J_{AT}$, “adding” j individuals in age-sex category A to each household with rank satisfying

$$C_{kAT} \sum_{i=0}^{j-1} \pi_{ikAT} < \text{rank} \leq C_{kAT} \sum_{i=0}^j \pi_{ikAT}$$

With this approach, the total number of extra individuals of age-sex category A allocated to non-CCS households of type T in the processing block who were recorded as having k individuals in age-sex category A at the Census is

$$E_{kAT} = C_{kAT} \sum_{j=1}^{J_{AT}} j \pi_{jkAT}$$

and the total number of extra individuals in age-sex category A imputed for non-CCS households of type T in the processing block is

$$E_{AT} = \sum_{k=0}^{L_{AT}} E_{kAT}$$

where L_{AT} denotes the maximum number of individuals in age-sex category A who were recorded in households of type T in the processing block at the Census. In situations where $L_{AT} > K_{AT}$, we put $E_{kAT} = 0$ for $k > K_{AT}$.

The total number $N_{AT}(\text{final})$ of all individuals (imputed + counted) in age-sex category A in households of type T in the processing block is then given by

$$N_{AT}(\text{final}) = \sum_{k=0}^{L_{AT}} k C_{kAT} + E_{AT} + n_{AT}(\text{CCS}).$$

To illustrate, suppose that $K_{AT} = 3$, $M_{AT} = 2$ and the values of $\{\pi_{jkAT}\}$ are

	j = 0	j = 1	j = 2
k = 0	0.99	0.01	0.00
k = 1	0.95	0.04	0.01
k = 2	0.90	0.08	0.02
k = 3	0.90	0.05	0.05

Furthermore, the values of C_{kAT} for this age-sex category and household type are

C_{0AT}	C_{1AT}	C_{2AT}	C_{3AT}
1400	1100	700	200

The total number of such individuals (imputed and actual) in non-CCS households of this type is then

	j = 0	j = 1	j = 2
k = 0	1386	14	0
k = 1	1045	44	11
k = 2	630	56	14
k = 3	180	10	10

Observe that the total number of such “non-CCS” individuals counted by the Census is 3100, but the total number of imputed + actual individuals is 3294. That is, an extra 194 age-sex A individuals have been imputed to the non-CCS households of type T.

An important point to note here is that there is no guarantee that the value of $N_{AT}(\text{final})$ above will agree with the ONC estimate $N_{AT}(\text{ONC})$ of the total number of individuals in age-sex category A in households of type T in the processing block. Put

$$I_{AT} = N_{AT}(\text{ONC}) - \sum_{k=0}^{L_{AT}} kC_{kAT} - n_{AT}(\text{CCS}).$$

Then the required condition for $N_{AT}(\text{final})$ to be “calibrated” to the ONC estimate is for the probabilities $\{\pi_{mkAT}\}$ to satisfy

$$I_{AT} = \sum_{k=1}^{K_{AT}} \sum_{j=1}^{J_{AT}} (jC_{kAT})\pi_{jkAT}$$

In general, this condition will not be satisfied. However, using well known results from calibration theory we can construct a set of probabilities that have this property and are as “close” as possible to the original probabilities $\{\pi_{jkAT}\}$ derived from the CCS. That is, given the “original” $\{\pi_{jkAT}\}$ we can define a “new” set of probabilities $\{\theta_{jkAT}\}$ such that the

“distance” between these two sets of probabilities is as small as possible and the “new” probabilities satisfy the constraints

$$I_{AT} = \sum_{k=0}^{K_{AT}} \sum_{j=0}^{J_{AT}} (jC_{kAT}) \theta_{jkAT}$$

and, for $k = 0, 1, \dots, K_{AT}$,

$$1 = \sum_{j=0}^{J_{AT}} \theta_{jkAT} .$$

The imputation method outlined above is then based on the $\{\theta_{jkAT}\}$, rather than the $\{\pi_{jkAT}\}$.

The development so far has assumed that all households of type T in the processing block are counted in the Census. In practice of course, the CCS will tend to find households that are missed by the Census. (Note we are still assuming that the CCS “count” of any characteristic is always at least as great as that recorded in the Census.) Let m_{kT} denote the number of households of type T and size k counted by the Census and h_{kT} the corresponding count of households missed by the Census but found by the CCS (in the CCS locations in the processing block). The probability that a household of type T and size k is missed by the Census is then

$$\phi_{kT} = \frac{h_{kT}}{m_{kT} + h_{kT}}$$

Given these probabilities, we can in theory “replicate” $R_{kT} = C_{kT}\phi_{kT}$ of the C_{kT} households of type T and size k in the non-CCS locations to account for missed households there. Precisely which households to replicate is a problem however. This is because households of the same size can have quite different age-sex compositions.

Let S be a variable denoting the age-sex “composition” of a household. Typically, S will be multidimensional, denoting the numbers of individuals in the household in each of $a = 1, \dots, A$ age-sex categories. The size of the household is then just the sum of all components of S. Given the CCS data, we can carry out a modelling exercise (e.g. using a multinomial logistic model) to obtain an expression for the probability that a household of type T and size k is missed by the Census which is a function of its age-sex composition. We denote this function by $\phi_{kT}(s)$, where s is a dummy variable which can take on all the observed values of S for households of size k and type T in the CCS locations. Replication of households in non-CCS locations can then be carried out conditionally on their age-sex composition. That is, for each “feasible” value of s we duplicate $R_{kT}(s) = C_{kT}\phi_{kT}(s)$ non-CCS households of type T, size k and age-sex composition s. This is easily accomplished by randomly selecting $R_{kT}(s)$ non-CCS households with these characteristics and then replicating them in the Census database.

As in the case of within household imputation explored earlier, the issue of “calibration” of the counts of individuals within age-sex categories within the database also arises here. One strategy for overcoming this problem, which should work provided most individuals are

imputed within counted households rather than in imputed households, is to first carry out the household imputation exercise described above and then, treating all non-CCS households (including imputed ones) as having been counted, carry out the individual within household calibrated imputation exercise described previously. That is, the counts C_{kAT} used there now reflect actual counted individuals as well as individuals from imputed households. Essentially, this treats the individuals created within imputed households as a “residual” which is then adjusted for in the final within household imputation step based on the calibrated probabilities $\{\theta_{jkAT}\}$.

4. Conclusion

The scheme outlined above has a number of limitations, most noticeably that its implied coverage probabilities for individuals are only a function of their age and sex and their household size and type. In addition, by effectively treating individuals within different age-sex groups independently in the imputation process, it does not make use of information from the CCS on the joint probability of noncoverage for individuals in different age-sex groups. On the other hand the preceding analysis does suggest that imputation methods can be developed that would preserve consistency between the household and the individual level in the hierarchy, although it should be emphasised that the ideas set out above are still subject to further discussion and change. Further thinking is required to refine the procedures, and more importantly a programme of work to evaluate the various methods is required.

April 1998