Advisory Group Paper (00) 13

EDITING AND IMPUTATION FOR THE 2001 CENSUS

1. This paper gives an overview of the editing and imputation system being developed for the 2001 Census. The edit rules and imputation methodology will be evaluated once the 1999 Census Rehearsal data has been run through the system and changes will be made where necessary for 2001.

2. The English Census form referred to in the paper has been published on the National Statistics web site at

   www.statistics.gov.uk/census2001/censusregs2000.asp

   as part of the Census regulations which were laid before Parliament on 6 June 2000.

3. **Advisory Group members are asked to:**
   - **Note the methodology planned for editing and imputation; and**
   - **Comment at the meeting**

Faith Anderson/Keith Whitfield
Census Division
Office for National Statistics

August 2000

Contents

EDITING AND IMPUTATION FOR THE 2001 CENSUS

1. Introduction

1.1 As with any data collection exercise, Census records can contain errors and missing values. The editing and imputation process is designed to correct obvious inconsistencies and to estimate values for missing data as accurately as possible and so as to preserve the relationships between variables. Additionally, for the 2001 Census we aim as far as possible to follow these principles:

- All changes that are made will improve the quality of the data
- The number of changes to inconsistent data are kept to a minimum
- As far as possible missing data will be imputed for all variables, so as to provide a complete and consistent database
- The system must be relatively easy to develop and be able to process large amounts of data within short timescales.

1.2 This paper discusses how the 2001 system will apply the following processes:

- **Multi-tick rules:** for cases where several answers are ticked instead of one
- **Range checks:** to deal with invalid answers
- **Filter rules:** where respondents disregard 'filter' instructions
- **The editing process:** to correct inconsistent responses between questions
- **Imputation:** to deal with missing data

1.3 This system is designed to fill the gaps in existing person and household records. A person is taken to exist if at least two of the name, date of birth and sex fields are completed. The One Number Census process imputes for whole households and people who were missed from the Census (see Annex E).

2. Multi-tick rules

2.1 These rules resolve multi-ticking of questions where only one box should have been ticked. In some cases there will be a rule for selecting one tick. If more than half the boxes have been ticked or we cannot decide on priorities for accepting one tick, the question will be treated as if the answer had been missing.

2.2 Annex A lists a summary of the multi-tick rules. These rules will be implemented during Data Capture.

3. Range checks

3.1 Answers which are outside an acceptable range will also be identified at the Data Capture stage and set to invalid. These are:

- Households:	with 0 or more than 99 rooms with more than 20 cars
- People:	with a date of birth before 1891 or after 29 April 2001 who last worked before 1941 working more than 99 hours per week

4. Filter rules

4.1 Filter rules are applied to resolve some inconsistencies and to decide which fields should be set to 'No Code Required' where questions were answered but should not have been. For example, children under 16 should not answer any of the employment questions. These rules will be applied at the Data Capture stage to avoid the cost of coding answers which would later be set to 'No Code Required'. The major outcomes of the filter rules are listed in Annex C.

5. The editing process

5.1 Editing identifies and resolves inconsistencies in the data using an edit ‘matrix’ in a similar way to 1991. This process takes place after Data Capture.

5.2 Hard checks will be carried out to identify inconsistencies which will not be permitted to remain in the data. In addition, there are soft checks to identify situations that may indicate mistakes but which may occur in a substantial minority of cases. The number of records failing these checks will be monitored with the intention of considering whether to revise the limits for edits in the 2001 Census. These checks are listed in Annex B.

5.3 The hard checks have been translated into a set of rules. Where possible, a variable is set to a particular value. Otherwise it is marked for imputation.

5.4 For people in households, the system deals with ‘within person’ inconsistencies first. There are three stages:

Stage 1: A set of rules dealing with within-person consistency checks involving age.

Stage 2: Rules to sort out other within-person inconsistencies, such as travel to work stated as ‘mainly at or from home’ conflicting with workplace address.

Stage 3: Resolution of between-person inconsistencies involving relationships. If a parent is less than 13 years older than a child, we will check whether the inverse relationship works, and similarly with grandparents/grandchildren. Otherwise the relationship fields are set to missing. As these between-person checks are done after the within-person checks involving age, we cannot guarantee that the minimum change principle will always be followed.

5.5 There are also rules to deal with inconsistencies in questions about the household as a whole and its accommodation. For people in communal establishments, there is a simpler set of rules as no relationship information is collected.

5.6 Other edit rules are applied as a result of the filter rules and the derivation of Activity Last Week (from Q17-21 on the English Census form), and also from the method of imputation. The edit rules are set out in Annex C.

6. Imputation

6.1 A variable may require imputation because it has been set to missing as a result of:

- No answer on the Census form
- The Data Capture system set the field to 'failed multi-tick' or 'invalid'
- The filter rules marked it for imputation
- The Editing process marked it for imputation to resolve an inconsistency

6.2 The process imputes values for household variables, people in households and people in communal establishments. The rest of this section deals with households with two or more people which is the most complicated case.

6.3 The principle of a Donor Imputation System is to search for a single donor household to supply all the missing variables in a recipient household. The search looks at all records in an Estimation Area, which is the same as the design group used for the One Number Census, ie a group of contiguous Local Authority Districts of about 500,000 population.

6.4 The method searches for a donor using up to five matching variables, which are determined by the fields requiring imputation on the recipient record. Values are copied over from the donor household to fill the missing values on the recipient record. The hard consistency checks are then applied and the donor is rejected if any check fails.

6.5 Potential donor households are scored using a second set of matching variables which relate to all people in the household. In addition, potential donors are penalised if they have been used before or if any of their fields have been edited or imputed. A record cannot be used as a donor if any of the fields to be imputed are also missing on the donor. If potential donors still score equally, the donor that is geographically closest to the recipient is chosen.

6.6 Ideally we would like to use a single donor household to impute values for all the people with missing values in a recipient household as this will preserve the joint distributions between variables. If we cannot find a suitable donor household for joint imputation we attempt to find donors to provide values for each separate person in the household, if necessary reducing the number of variables we match on. Further details of the imputation system are at Annex D.

6.7 For the Census Rehearsal, this is the final stage of the main imputation. We will assess the number of records not finding a donor and depending on the scale of the problem assess what our final fall-back process should be for 2001. However, there may be two additional imputation stages:

1) a 'mop-up' relationship imputation to deal with records left with missing relationship data

2) a process to impute postcode of workplace (or travel address in Scotland) and postcode of usual address one year ago.

Research into both these processes is currently being carried out.

6.8 We are proposing to impute the full detail of the code for occupation, industry, Country of Birth and ethnic group. The form of the ethnic group question varies between the England and Wales, Scotland and Northern Ireland forms.

7. Evaluation

7.1 The 2001 Donor Imputation System (DIS) has been evaluated using 1991 Census data and has been compared with the 1991 hot-deck system and with a prototype solution based on Neural Networks. The approach consisted of the following stages:

- Use two complete 1991 Census LAD datasets
- Create missing values in these datasets at random
- Run each system against the datasets
- Compare the data of the imputed and real datasets to assess:
    - whether or not any inconsistent responses were imputed
    - how often the correct value is imputed; and
    - how well the marginal and joint distributions are preserved.

7.2 The results showed that on all criteria the DIS out-performed the hot-deck method. The Neural Network approach did not perform well and in some instances imputed inconsistent data. This research is described in more detail in Vickers & Yar (*Proceedings of the Joint IASS/IAOS Conference, 1998*) and Cruddas, Thomas & Chambers (*1997 Statistics Canada Symposium*) – available on request.

7.3 Further evaluation of the edit and imputation system is planned using the 1999 Rehearsal data. This will review:

- each consistency check rule
- the multi-tick rules
- the filter rules
- the matching variables for imputation; and
- the quality of imputation for all variables but in particular for those questions that have changed since 1991 or are new for 2001

7.4 For the consistency checks, multi-tick and filter rules the approach will consist of looking at the incidence of each rule being invoked to see whether or not the rule produced the correct value.

7.5 For imputation a similar approach will be adopted to that for the evaluation mentioned above. We will also assess to what extent any bias has been removed from the data by the imputation process (this has already been done for a selection of variables).

8. Comparison with 1991 Census

8.1 The final 2001 Census database will contain complete data for all households and individuals where appropriate. Some records will contain one or more values which have been imputed through the editing and imputation suite; others will have been completely imputed through the One Number Census process.

8.2 The editing of inconsistent data will be carried out in a similar manner to 1991 using an edit matrix, with the aim of minimising the number of changes.

8.3 Edit elimination tables were set up for 1991 because they were restricted to the easy to code items which were processed for all forms. In 2001, all questions will be 100% processed. It would be impracticable to expand the edit elimination tables to cater for all items. Some of the consistency checks carried out in 1991 on items such as occupation will not be repeated in 2001 as there are only a small number of inconsistent combinations and the resources required would be disproportionate. A few 16 year old doctors or coal miners working in London may therefore appear in the output.

8.4 The imputation system will be different for 2001 in three ways:

- Imputation was only applied to the easy to code questions in 1991 but will be carried out for almost all variables in 2001.

- The missing values of a record are to be imputed as far as possible from a single donor in order to maintain relationships between variables. In 1991 each missing variable was imputed separately so that a record with several missing values was likely to have had more than one donor.

- The 2001 system considers all records in the Estimation Area as potential donors. The 1991 hot-deck system of imputation considered only those records which had already been processed, and did not search for donors among the subsequent records.

MULTI-TICK RULES ANNEX A

These rules were devised using evidence from the 1997 test. If more than half the tick boxes are ticked then usually no attempt is made to code. Otherwise the following rules apply. They will be evaluated after analysis of the Rehearsal data.

The level of multi-ticking in the 1999 Census Rehearsal was low. It was over 1% in some areas for ethnic group but this is covered by the coding rules. For travel to work it can be as much as 1% (mainly bus/on foot and passenger in car/on foot) but for other variables it was mostly between 0.1% and 0.2%.

HOUSEHOLD VARIABLES

H1	Type of accommodation	Accept last tick
H5	Lowest floor level	Accept first tick (ie the lowest floor ticked)
H7	Cars and vans	Accept last tick, and code any written-in value for four or more vehicles
H8	Owns/rents	'Owns outright' and 'owns with a mortgage' – code 'owns with a mortgage'
		'Owns with a mortgage' and 'part rent/part mortgage' - code ‘owns with a mortgage'
		'Rents' and 'rent free' - code 'rents'
		Otherwise, fails multi-tick
H9	Landlord	One of 'private landlord', 'employer of household member', 'relative' with 'other' - delete 'other'
		'Private landlord' and 'employer' - code 'employer'
		'Private landlord' or 'employer' with 'relative' – code 'relative'

PERSON VARIABLES

4	Marital status	Accept tick in priority order: separated, re-married, divorced, widowed, married, single.
7	Country of birth	Accept tick relating to country of enumeration. Otherwise accept first tick.
10	Health	'Good' and 'Fairly good' - code 'Fairly good' 'Fairly good' and 'Not good' - code 'Not good' Otherwise fails multi-tick.
24	Employee/ self-employed	'Self-employed with employees' and 'self-employed without employees' – code 'self-employed with employees'
25	Workplace size	Accept first tick, ie the lowest workplace size ticked
32	Workplace address	Where two ticks and no text, accept tick in priority order 'offshore installation', 'at or from home', 'no fixed place'
33	Travel to work	Accept tick in priority order (England & Wales order): at or from home, train, underground, bus, passenger in car, driving car, motor cycle, taxi, bicycle, on foot, other

Person Hard Checks (within person)

1. A child under 5 cannot provide substantial unpaid personal help.
2. A person with travel to work ‘mainly at or from home’ must have workplace address ‘mainly work at or from home’ and vice versa (not Scotland).
3. If a person is self-employed without employees, number of people employed should be 1-9.
4. A child under 16 must have marital status of single, unless their Country of Birth is ‘elsewhere’. Country of Birth ‘missing’ should also fail the check.
5. A child aged 6-15 must be a schoolchild/student.
6. If age is 16-74, Activity Last Week (ALW) cannot be ‘No Code Required’ and for other ages must be ‘No Code Required’.
7. A person answering ‘No’ to Q5 (schoolchild/student) cannot have ALW ‘economically inactive – student’.
8. A person answering ‘Yes’ to Q5 cannot have ALW ‘economically inactive – retired’, ‘looking after home/family’, ‘permanently sick’ or ‘other’.

Note that some of these checks, eg check 6, result from the actions taken at the filter rule stage. There are other checks to prevent imputation of values that contradict the filter rules.

Additional checks for Scotland

1. A schoolchild/student cannot have ‘not currently working or studying’ for method of travel.
2. A schoolchild/student cannot have ‘not currently working or studying’ as their address travelled to.
3. If method of travel is ‘not currently working or studying’, address travelled to must also be ‘not currently working or studying’ and vice versa.
4. If method of travel is ‘work or study mainly at home’, then address travelled to must also be ‘work or study mainly at home’ and vice versa.
5. If method of travel is ‘not currently working or studying’, the person could not be in a job last week and cannot be a ‘student’.
6. If address travelled to is ‘not currently working or studying’, the person could not be in a job last week and cannot be a ‘student’.
7. If a person aged 16-74 has a method of travel or travel address other than ‘not currently working or studying’ or ‘No Code Required’, then ALW must be ‘working’ or ‘economically inactive – student’.
8. If travel address is ‘offshore installation’ then job last week should be ‘Yes’.

Person Soft Checks (within person)

1. Children under 16 and with Country of Birth ‘elsewhere’ are unlikely to have marital status other than single.
2. People aged 16 or 17 are unlikely to be divorced.
3. People under 35 are unlikely to be ‘retired from paid work’.
4. People aged 55 and over are unlikely to be a student.

Additional checks for Wales, Scotland, and Northern Ireland

5. A child under 2 is unlikely to be able to speak Welsh/Gaelic/Irish.
6. A child under 3 is unlikely to be able to read or write Welsh/Gaelic/Irish.

Relationship Hard Checks

1. A person who has a husband/wife in the household must have marital status of married, remarried or separated.
2. A husband and wife must be of opposite sex.
3. Parents of the same person must be of opposite sex.
4. Children with at least one parent in common cannot be spouse/partner of each other.
5. A person who has a partner living in the household must be aged 16 or over.
6. A parent must be 13 or more years older than their child.
7. A grandparent must be at least 26 years older than their grandchild.
8. A person can only have one partner or husband/wife in the household.
9. A person can only have a maximum of two parents (excludes step-parents).

Relationship Soft Checks

1. A parent is unlikely to be only 13 or 14 years older than their child.
2. A grandparent is unlikely to be 26 to 29 years older than their grandchild.
3. Two people living as partners are unlikely to be of the same sex.
4. A stepchild is unlikely to be older than his stepfather/mother.
5. Brothers and sisters are unlikely to have an age difference greater than 30 years.
6. It is unlikely that a person would have more than one step-parent in a household.
7. It is unlikely that a person would have more than two mother/fathers and stepmothers/fathers in the same household.
8. A step-parent is unlikely to be under 16.
9. It is unlikely that a mother will be 50 years older than her son/daughter.
10. The oldest person in the household is unlikely to be less than 16.

Household Hard Checks

1. Caravans and other mobile and temporary structures cannot have more than 10 rooms.
2. The ‘lowest floor level of the household’s living accommodation’ for a caravan, mobile or temporary structure can only be ‘ground’ or ‘first’.
3. The ‘lowest floor level of the household’s living accommodation’ for a household with building type of ‘whole house or bungalow’ can be no higher than ‘first’.
4. A household’s accommodation cannot be self-contained if there is no sole use of bath/WC.
5. A household not living in self-contained accommodation must have building type ‘part of converted or shared house’.

Household Soft Checks

1. Caravans and other mobile or temporary structures are unlikely to be rented from the Council (local authority), Scottish Homes, Housing Association or Charitable Trust.
2. Caravans and other mobile or temporary structures are unlikely to have central heating.
3. Accommodation rented from the council is unlikely to be rent free.

EDIT RULES ANNEX C

PERSON RULES

Rules resulting from filter rules

- If age is 6-15 and schoolchild/student is ‘No’ or missing, set schoolchild/student to ‘Yes’
- If there are responses to any of Q7 onwards (ie after term-time address) and schoolchild/student is missing, set schoolchild/student to ‘No’
- If age is under 16, year of birth is not 2001 and Q15-34 (qualifications and employment) are answered, accept age and change the other answers to ‘No Code Required’. If year of birth is stated as 2001 and Q15-34 are answered, set age to Missing.
- If the answer to Q17 (Worked last week) is ‘Yes’ but all the subsequent economic activity questions contradict this, change Q17 to 'No'.
- *(Additional rule for Scotland)* If age is 16-74 and travel address is ‘No’ or missing, ALW will be imputed in the range ‘not working’.

In most circumstances date of birth is taken as accurate when it is inconsistent with other answers.

Rules arising from the check between Activity Last Week (ALW) and whether schoolchild/student:

- If ‘No’ to Q5 (schoolchild/student) and ALW is ‘economically inactive – student’, change ALW to ‘economically inactive – other’.
- If ‘Yes’ to Q5 and ALW is ‘economically inactive – retired’, ‘looking after home’, ‘permanently sick’, or ‘other’, change ALW to ‘economically inactive – student’.

In both cases the answer to Q5 (which may have been changed as a result of the filter rules) is considered to be more reliable than the answer to Q21. This will be evaluated after the Census Rehearsal.

EDIT PROCESS

Stage 1

This looks at age, marital status, country of birth (CoB), carer and relationship (whether or not the person has a spouse/partner in the household). There are actions for each combination of these variables, which have been decided either by following the principle of minimum change or, where this does not dictate which variable to change, by applying the following rules:

- If carer is missing, set to ‘No’ unless ALW is also missing, in which case leave carer as missing and impute with ALW.

- If CoB is missing for a person in a household with two or more people then :

  if siblings are present in the household and they all have the same CoB code, or there is only one sibling, assign the CoB code of the sibling(s);

  or if parents are present in the household and they both have the same CoB code, or there is only one parent, assign the CoB code of the parent(s);

  or if other related persons are present in the household and they all have the same CoB code, or there is only one other related person, assign the CoB code of the other related person(s)
- If relationship conflicts with age or marital status, change relationship (eg if age is 20, marital status is 'single', carer is 'Yes', spouse is 'Yes' and partner is 'No' change the relationship that is spouse to missing)
- If carer conflicts with age, change carer.
- If marital status conflicts with age, change marital status.

'Minimum change' may involve making a change to the value of a person's age. In some cases we will then change all the economic activity questions from 'No Code Required' to missing. For example, if age is 4, marital status is married, the person is a carer and has a spouse then age would be corrected. However, the filter rules will have set the economic activity questions to 'No Code Required'. Unless economic activity is changed, the consistency checks will mean that no age in the range 16-74 can be imputed. We will evaluate this after the Census Rehearsal.

Stage 2 (Rules marked * do not apply to Scotland)

This deals with the other 'within person' inconsistencies and some missing values:

- * If travel to work is 'mainly at or from home' and workplace address is missing, set workplace to 'mainly work at or from home'
- * If travel to work is missing and workplace address is 'mainly work at or from home', set travel to work to 'mainly at or from home'
- * If travel to work is 'mainly at or from home' and workplace address is not 'mainly work at or from home', set travel to work to 'not stated'.
- * If workplace address is 'mainly work at or from home' and travel to work is not 'mainly at or from home', set travel to work to 'mainly at or from home'.

The last two rules mean that workplace address is taken as correct when it conflicts with mode of travel to work.

- If person is 'self-employed without employees', set number of people working for employer to '1-9'.
- If health is missing, set to 'Good', unless ALW is also missing in which case leave health as missing and impute with ALW.
- If supervisor is missing, set to 'No' unless occupation is also missing in which case leave supervisor as missing and impute with occupation.

Rules for Scotland

For Scotland the additional consistency checks involving age, schoolchild/ student, method of travel, address travelled to and activity last week have not been translated into rules because the number of possible combinations is too great.

Instead, we have agreed the following procedure to determine, where possible, the ‘minimum change’ action:

> Apply the ten consistency checks. Count how many checks fail for each variable. Set the variable with the highest count to missing (if two or more are highest, set both or all of them to missing).
>
> Apply the ten checks again. Count how many checks fail for each variable. Set the variable(s) with the highest count to missing.
>
> Continue until no checks fail.

Stage 3 (Between person)

- If a husband/wife are not of the opposite sex, set relationship to missing (ie mark for imputation).
- If a person has a spouse and a partner, choose the spouse in preference (ie set the ‘partner’ relationship to missing)
- If a person has more than one spouse or more than one partner, set both relationships to missing.
- If parents of the same person are not of opposite sex, set both parent/child relationships to missing.
- If a person has more than two parents, set all the parent/child relationships to missing.
- If two children of a parent are spouse/partner of each other set the spouse/partner relationship to missing.
- If a parent is not at least 13 years older than their child, change the relationship to child/parent. If there is now no inconsistency accept this. Otherwise believe age and set relationship to missing.
- If a grandparent is not at least 26 years older than their grandchild, change the relationship to grandchild/grandparent. If there is now no age inconsistency accept this. Otherwise set relationship to missing.

The last two rules are designed for cases where reciprocal relationships have been entered. However we need to be careful that we cover implicit edits to avoid searching for a donor for households containing, say, a 12 year old parent. So we need the following additional rules:

- Where the age of the ‘child’ is missing in a parent/child relationship, if the ‘parent’ is under 13 set relationship to missing.

- Where the age of the ‘child’ is missing in a parent/child relationship and the ‘parent’ is aged 28 or less, if the marital status of the ‘child’ is anything other than single then set relationship to missing unless Country of Birth is ‘elsewhere’.
- Where the age of the ‘grandchild’ is missing in a grandparent/grandchild relationship, if the ‘grandparent’ is under 26 set relationship to missing.
- Where the age of the ‘grandchild’ is missing in a grandparent/grandchild relationship and the ‘grandparent’ is aged 41 or less, if the marital status of the ‘grandchild’ is anything other than single then set relationship to missing unless Country of Birth is ‘elsewhere’.

Communal persons

- If residential classification is missing, set to 'other'.

HOUSEHOLD RULES

Rules resulting from the filter rules

- If answer to H8 (own/rent) is ‘owns’ but H9 (landlord) has been answered as well, change H9 to ‘No Code Required’.
- If H9 is answered but H8 is missing or multi-ticked, set H8 to ‘rents’.

Edit process

Stage 1

This looks at type of accommodation, whether self-contained and sole/shared use of bath/shower and toilet. In some cases, minimum change dictates which field to change and in other cases the following rules dictate the change:

- If H1 (type of accommodation) is other than ‘part of converted or shared house’ and H2 (self-contained) is missing, set self-contained to ‘Yes’.
- If H1 is other than ‘part of converted or shared house’ and H4 (bath/shower) is missing, set bath/shower to ‘Yes’.
- If there is a conflict between self-contained and bath/shower, change bath/shower.

  Example: H1 is ‘part of converted or shared house’, self-contained is ‘Yes’ and bath/shower is ‘No’ - change bath/shower to ‘Yes’.
- If bath/shower is ‘No’ and self-contained is missing, set self-contained to ‘No’.
- If there is a conflict between type of accommodation and bath/shower, change bath/shower.

  Example: H1 is ‘detached house or bungalow’, self-contained is missing and bath/shower is ‘No’: change bath/shower to ‘Yes’ and set self-contained to ‘Yes’.
- If there is a conflict between type of accommodation and self-contained, change self-contained.

  Example: H1 is ‘detached house or bungalow’, self-contained is ‘No’ and bath/ shower is missing: change self-contained to ‘Yes’ and set bath/shower to ‘Yes’.
- If type of accommodation and self-contained are both missing and bath/shower is ‘No’, set bath/shower to missing and impute all three fields.
- If type of accommodation is missing, self-contained is ‘No’ and bath/shower is ‘Yes’, change self-contained to missing and impute.

Stage 2

- If there is a conflict between rooms and type of accommodation, change rooms.
- If there is a conflict between floor level and type of accommodation, change floor level.

DONOR IMPUTATION ANNEX D

The Donor Imputation System (DIS) provides values for variables which require imputation for one of the following reasons:

- No information was provided for that variable on the Census form
- The Data Capture system set the variable to ‘failed multi-tick’ or ‘invalid’
- The filter rules marked the variable for imputation
- The Edit process marked the variable for imputation to resolve an inconsistency.

The system processes households, private persons and communal persons.

Households are processed in the following order:

1. Only one variable missing
2. Only one person in the household with variables missing
3. Two or more person records in error in the household

This increases the pool of donors for the more complicated cases.

Imputation for people in households

The following steps are carried out in the order listed when attempting to impute values for people within households:

1. Joint imputation (not needed for one-person households).
2. Individual record imputation, matching on household size.
3. Individual record imputation, not matching on household size.
4. Individual record imputation, not matching on household size and using a gradually reduced set of PMVs.
5. Final fall-back process. For Rehearsal, this is to report the records which have failed imputation.

Imputation for people in communal establishments

1. Individual record imputation.
2. Report records failing imputation.

For household variables

1. Individual record imputation, using a household record to impute all missing household variables in a recipient household.
2. Report records failing imputation.

The Matching Process – Finding a Donor

The search for donors takes place within EDs with the same hard to count index, although this condition may be relaxed after evaluating Rehearsal data. There are two stages to the matching process, with records failing to find a donor in the first stage progressing onto the second. Records that fail to find a donor at the second stage are recorded as failing imputation.

First stage:

- Match on PMVs and household size. A perfect match must be achieved on the PMVs. Choose the best donor by scoring other people in household on Secondary Matching Variables (SMVs). The search is stopped if a perfect match is found on the SMVs.

- Where there are donors with equally good matches the one nearest to the recipient (lowest value of DISTGEO) is chosen.

Second stage:

- Match on PMVs but not household size. A perfect match must be achieved on the PMVs. Choose the first donor found. No scoring on SMVs.

Primary Matching Variables

Each variable has a set of Primary Matching Variables (PMVs). The PMVs were those found to be most highly associated with the variable that is missing. That is, matching on the chosen PMVs gives the highest chance of imputing the 'correct' missing response, and preserving the marginal distributions within the data.

If more than one variable requires imputation, the PMVs for each of these variables are combined. There will be a maximum of five for this combined set of PMVs. If more than a certain number of variables require imputation, the PMVs are the four highest importance variables which have a value. Where there are fewer than four variables with a value, the available ones are used.

For household variables the number of PMVs is limited to one or two on the assumption that in most cases the best donor household will be the one next door. More importance is therefore placed on DISTGEO. The PMVs are used to prevent situations where the nearest neighbour is a different house type.

The PMVs for all communal person variables are the same as those for the corresponding private person variable, but with the addition of Position in Establishment and deletion of Relationship where appropriate.

Grouping of PMVs

The accuracy of this approach was examined using the Imputation Testing Program. After testing all variables that contain ‘groupable’ PMVs it was generally found that there is no loss in imputation accuracy when using the grouped PMVs. In some cases the quality of imputation was actually found to improve.

Determination of the PMVs

The PMVs have been determined by manipulating the output from the CHAID package, a classification and regression approach for nominal variables based upon chi-squared significance tests between the response variable and its best predictors. The analysis is based on 1991 Census data. The PMVs were modified to reflect the change to a resident population base and the different filters in the 1999 form. We also applied some common sense.

Examples of PMVs are, in priority order:

For **Marital Status:** Relationship to Person 1, Age, Sex, Highest Qualification
For **Sex:** ALW, Relationship to Person 1, Marital Status, Occupation
For **ALW:** Age, Ever Worked, Sex, Limiting Long Term Illness

If Marital Status, Sex and ALW are all missing on the recipient person, the combined PMVs will be:
Relationship to Person 1, Age, Ever Worked, Highest Qualification, Occupation

Secondary Matching Variables

In most cases there will be a standard set of Secondary Matching Variables (SMVs): *Relationship to person 1, Age, Marital Status* and *Sex* regardless of the variable being imputed. The SMVs refer to all other members of a household, and are used to distinguish between donors with equally good matches on PMVs. Thus the selected donor has as similar as possible a 'family' structure, so as to preserve the joint distributions between variables.

There are exceptions when the additional variables actually provide information about the missing response. Thus the variables *Ethnic group, Country of Birth, Language, Usual Address One Year Ago* will include the additional SMVs *Ethnic group, Country of Birth, Language, Usual Address One Year Ago* respectively.

The Statistical Distance (DISTSEC) is calculated by comparing the values for the SMVs on the corresponding person records. Each variable that matches scores zero; each variable that is different adds 1 to the score. The search for a donor is stopped if a perfect match is found on the SMVs. Where donors have equally good matches the one closest to the recipient is selected.

Using a Standard Set of SMVs

We tested several variables to examine if a standard set of SMVs can be used for all variables. It is desirable to use a standard set of SMVs if possible to simplify the imputation system. The standard set consists of *Relationship to Person 1, Marital Status, Age grouped in 5 year bands, Sex.*

The analysis was completed on two person households only, which would be expected to provide a worse set of results than larger households.

Except for *Ethnic Group and Country of Birth*, the results suggest that it is acceptable to progress with a standard set of SMVs. It seems necessary to match on the *Ethnic Group/Country of Birth* of all household members when one person has their *Ethnic Group/Country of Birth* missing.

For the 1999 Census Rehearsal, similar treatment will also be afforded to the variables *Language* and *Usual Address One Year Ago*, as we would expect the responses of other household members to have a direct relationship with the missing persons for these variables. All rules need to be consistent regardless of household size, and we therefore continue to examine larger households to confirm that this outlined approach is suitable. It will be examined after the Rehearsal data are analysed.

Further Analysis – Three and Four Person Households

Research into three-person and four-person households broadly confirmed the above findings, and we are therefore confident that the SMV analysis should be formulated in the outlined manner.

Geographical distance – DISTGEO

This is the geographical distance between the recipient and donor households (or communal establishments), and is used to distinguish between donors with equally good matches on PMVs and SMVs. It is calculated from household grid references, which are based either on the address, postcode centroid or ED centroid.

Weights

We wish to apply penalties to records when a record has been used as a donor, when it has had data changed as part of the edit process, or when it has had data changed as part of the imputation process.

However, these penalties should not be equal and might be assigned as follows:

A penalty of 1 to a record if it has already been used as a donor.
A penalty of 2 to a record if any field has been changed as a result of an edit.
A penalty of 3 to a record if any field on it is imputed.

For the Rehearsal we will use equal weights under all three circumstances, but varying weights will be assigned for 2001 if appropriate.

Who cannot be used as a donor

A person record cannot be a donor if it is in the same household as the recipient. A record cannot be a donor if the fields to be imputed are currently missing on the potential donor.

Households with more than one person in error

We carry out a search throughout the Estimation Area, matching on Household size and the Hard to Count Index, for donor households with a perfect match on the PMVs for each person in error in the household.

DISTSEC is used to choose between potential donors. The remaining people in both the recipient and donor households are ordered by age, sex and marital status. DISTSEC is calculated for each person in the household, whether or not they are in error. There are no thresholds on the DISTSEC score so it does not matter if fields that have already been used as PMVs are double-counted as SMVs.

The search for donors can stop if a donor is found with DISTSEC=0 and is in the same Local Authority district as the recipient.

If no donor is found at joint imputation, we search through the Estimation Area for a donor for each person in error in turn, matching on household size and the PMVs for that person.

We choose between potential donors using DISTSEC, and order the remaining persons in both the recipient and donor households by age, sex and marital status. We calculate DISTSEC for each other person in the household and stop the search if DISTSEC=0 and the donor is in the same LA district as the recipient.

If any person records are still in error, we search for a donor for each person in turn, matching on the PMVs for that person. To decide between donors we use DISTGEO rather than DISTSEC which cannot meaningfully be compared for households of different sizes.

If a person record is still in error, we search for a donor by reducing the number of PMVs in a predefined order until only matching on one PMV. If no donor can be found, a list of records not imputed will be output for the Census Rehearsal.

Note that at any stage a donor will be rejected if the imputed values from it create inconsistencies in the recipient household which would fail the hard checks described in Annex B.

Household Variables to be Imputed

QUESTION	CATEGORIES	LEVEL OF IMPUTATION
Type of Accommodation	1 Detached 2 Semi-detached 3 Terraced 4 Purpose built flats 5 Part of converted or shared house 6 Commercial building 7 Caravan or other mobile structure	Impute codes 1-7
Self-contained	1 Yes 2 No	Impute codes 1-2
Rooms	01-99, NCR	Impute rooms 01-99
Bath/shower	1 Yes 2 No	Impute codes 1-2
Lowest Level of Accommodation	1 Basement 2 Ground 3 First 4 Second 5 Third or fourth 6 Fifth or higher	Impute codes 1-6
Central heating	1 Yes 2 No	Impute codes 1-2
Cars	00 None 01-09 1-9 10 10-20	Impute codes 00-10
Tenure	1 Owns outright 2 Owns with a mortgage 3 Part rent and part mortgage 4 Rents 5 Rent free	If Landlord is answered and Tenure is missing, Tenure will be set to 'Rents' in the filter rules. Otherwise impute codes 1-5
Landlord	1 Council 2 Housing association 3 Private landlord 4 Employer 5 Relative or friend 6 Other	Impute codes 1-6
Furnished/ unfurnished (Scotland only)	1 Furnished 2 Unfurnished	Impute codes 1-2
Living area on more than one floor (N. Ireland only)	1 More than one floor 2 One floor NCR	Impute codes 1-2

NCR – No code required

Person Variables to be Imputed

QUESTION	CATEGORIES		LEVEL OF IMPUTATION
Relationship	1 Husband/wife 2 Partner 3 Son/daughter 4 Step-child 5 Brother/sister 6 Mother/father	7 Step-mother/ step-father 8 Grandchild 9 Grandparent 10 Other related* 11 Unrelated* NCR	Impute codes 1-11 for all columns
Sex	1 Male	2 Female	Impute codes 1-2
Age	0-110		Impute age 0-110
Marital status	1 Single 2 Married 3 Re-married	4 Separated 5 Divorced 6 Widowed	Impute codes 1-6
Schoolchild/ student	1 Yes 2 No		Normally covered by filter rules and derivation of Activity Last Week. If key fields are missing, filter rules will not be able to set it, so it will be imputed
Term time address indicator	1 Yes 2 No		Normally covered by filter rules and derivation of ALW. If key fields are missing, filter rules will not be able to set it, so it will be imputed
Country of Birth	Codes 001-937, NCR		If CoB is missing for a person in a 2+ person household the edit rule on p.12 is applied. Full CoB code is imputed if the person is in a single person household, or has no related people in the household with the stated CoB, or the CoB codes for the siblings or the parents or the other related people are different, or lives in a communal establishment.
Ethnic	Codes 101-985, NCR		Impute full codes.
Language (not England)	1 Understand 2 Speak 3 Read	4 Write 5 None	Impute codes 1-5.
Health	1 Good 2 Fairly good	3 Not good	Imputed if ALW is also missing, otherwise Edit rule sets Health to ‘Good’ if missing
Carer	1 1-19 hours 2 20-49	3 50+ 4 No	Imputed if ALW is also missing, otherwise Edit rule sets to ‘No’ if missing
Limiting long term illness	1 Yes 2 No		Impute codes 1-2.
Usual address one year ago	1 Address on front of form 2 No usual address one year ago 3 Same as person 1 4 Elsewhere with postcode or country code		Imputation system under review
Qualifications	01 –15 NCR		All Qualification ticks from the donor record will be imputed to the recipient.
Activity Last Week (ALW)	01 Working 02 Working full-time 03 Working part-time 04 On a government training scheme 05 Available for work in next two weeks 06 Waiting to start a job 07 Ec. inactive – Retired		The derived ALW field will be imputed, not the answers to the component questions.

	08 Ec. inactive – student 09 Ec. inactive – looking after home/ family 10 Ec. inactive – permanently sick 11 Ec. inactive – other NCR	
Ever worked	1 Yes 2 No	Set by filter rules unless ALW is also missing.
Year last worked	1941-2001	Impute year 1941-2001.
Employment status	1 Employee 2 Self employed with employees 3 Self employed without employees	Impute codes 1-3.
Company size	1 1-9 3 25-499 2 10-24 4 500+	Edit rules set to 1-9 if employment status is 3. Otherwise impute codes 1-4.
Occupation	Codes 101-999, NCR	Impute full code
Supervisor	1 Yes 2 No	Only imputed if occupation is also missing, otherwise set to 'No' in edit rules
Industry	Codes 02-8514, NCR	Impute full code
Workplace Indicator (Travel Address in Scotland)	1 Mainly work* at or from home 2 Offshore installation 3 No fixed place 4 Postcode or country code * Work or study in Scotland. Scotland has additional category**:** Not currently working or studying	Imputation system under review
Transport to work (Includes transport to place of study in Scotland)	01 Work mainly at or from home 02 Underground etc 03 Train 04 Bus etc 05 Motor cycle etc 06 Driving a car or van 07 Passenger in a car or van 08 Taxi 09 Bicycle 10 On foot 11 Other 12 Not currently working/studying (Scotland) 13 Car or van pool sharing driving (NI)	Edit rules set to 01 if workplace address is 'mainly work at or from home'. Otherwise impute codes 01-11.
Hours worked	01– 99, NCR	Impute hours worked 01-99.

IMPUTATION IN THE ONE NUMBER CENSUS ANNEX E

An important aim of the One Number Census (ONC) is to allow the creation of a single, person level database adjusted for undercount. This database will then be used to generate all statistical output from the Census. In the past this database has reflected those individuals actually enumerated by the Census, although in 1991 provision was made for imputing households identified by enumerators who did not return a census form. However, the information on the characteristics of missed persons obtained in the Census Coverage Survey (CCS) will allow the creation of a database which represents our best estimate of the entire population, whether counted by the Census or not. This will be accomplished by a process of imputation, where additional individuals will be added to the Census database to account for those missed by the Census.

In household surveys, individual records are often given a weight to compensate for non-response. In producing tables, each record is multiplied by its weight before it is added to the relevant total. This procedure was considered for this final stage of the ONC but imputation is strongly favoured, provided it can be done satisfactorily.

Imputation is already accepted as standard practice where questions are left unanswered or are found to be invalid. Extending imputation to whole households and people is nonetheless a big step. A method of doing this has now been developed and is being tested using the 1999 Census Rehearsal data.

The process of matching the CCS and Census data will allow the characteristics of those households and individuals missed by the Census to be identified. It is expected that the characteristics of people within entirely missed households will differ from those missed from within otherwise counted households. However, once these features have been identified, prediction of both numbers and characteristics of missed individuals in the rest of the population not covered by the CCS will be possible.

The process of imputation is complex, reflecting the variability in the characteristics of those people found by the CCS to have been missed by the Census. However, it can essentially be broken down into three stages.

Stage 1 – Imputation of missed households

The first stage of the process imputes individuals in missed households on to the database. This is carried out by allocating a weight to every household counted by the Census corresponding to its propensity to have been missed by the Census. These weights are derived from an analysis of missed households in the matched CCS/Census data. Households with high weights (and hence likely to have been missed) are duplicated on the Census database using a systematic procedure which spreads these duplications over areas where missed households are most likely. These duplicated households are referred to as synthetic households, with the individuals they contain referred to as synthetic individuals.

Stage 2 – Imputation of missed individuals

The second stage of the process focuses on individuals who were missed in households actually counted by the Census. A weight is created for each individual (again based on information obtained from analysis of the matched CCS/Census data) which reflects their propensity to have been omitted from the census return for their household. These weights are then used to carry out a second systematic imputation of extra synthetic individuals into these households.

Stage 3 – Calibration to estimates of the population

A crucial requirement of the imputation process is that the overall distribution of synthetic individuals and households created by the above imputation process should be equal to the ONC estimates of the actual distributions of households and individuals missed in the 2001 Census. This calibration is accomplished by adjusting household and individuals weights appropriately in the imputation process, and by a final stage in the process, which either removes excess synthetic individuals and synthetic households from the Census database or tops up the database where necessary to ensure consistency with the ONC estimates. Removing or adding individuals within households has a knock on effect, changing a large number of distributions at household level as well as individual level in the database. However, this imputation process changes no individual level data actually collected in the Census.

Eventually, an individual level database will be created which will represent the best estimate of what would have been collected had the 2001 Census not been subject to underenumeration. Tabulations derived from this database will automatically include compensation for underenumeration. All ONC counts will be based on this database which includes imputed underenumeration.