

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Data Editing**  
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

**AN OVERVIEW OF EDIT AND IMPUTATION IN THE 2001 UK CENSUS**

Submitted by the Office for National Statistics<sup>1</sup>

**Contributed paper**

**I. INTRODUCTION**

1. The 2001 UK Census will take place on 29 April 2001 and will be conducted by the Office for National Statistics (ONS) for England and Wales, the Registrar General (Scotland) (GRO(S)) and the Northern Ireland Statistical and Research Agency (NISRA). It will be the 20<sup>th</sup> decennial Census to take place, the first being conducted in 1801 for Great Britain.

2. Census data, as with data from any data collection exercise can contain errors and missing values. The aim of any Edit and Imputation system is to correct these errors so they match as closely as possible the correct data and so that relationships between variables are preserved. Fellegi and Holt (1976) suggest that when more than one field is in error, the smallest number of fields should be changed in order to correct the data and that for imputation a joint imputation method should be used to impute missing data. Vickers and Yar (1998) describe the Donor Imputation System that will be used by the UK Census Offices to impute missing data in the 2001 Census. This paper will examine how inconsistencies will be resolved. Two processes will be used. The first uses a set of edit rules to fix errors where it is obvious which variable is causing the inconsistency and the value that this variable should take. The second process sets all variables involved in the inconsistency to missing and uses the donor imputation system to find a donor that will resolve the inconsistency. Both attempt to keep to the principle of minimum change but this is not guaranteed. Census Rehearsal data will be used to estimate how well the system meets the minimum change principal (the 1999 Census Rehearsal took place in April 1999 and this analysis will take place over the next year).

**II. ERRORS IN CENSUS DATA**

3. There are several ways that error can be introduced into Census data. These are outlined below.

**Multi-tick Responses**

4. In 2001 the UK Census will make much more use of tick boxes than previous censuses. For some questions it will be legitimate to provide more than one tick as a response, such as the qualification question.

---

<sup>1</sup> Prepared by Paul Vickers.

However, the majority of questions require only single tick responses. Clearly errors will be introduced if a respondent ticks more than one box or the optical mark recognition system picks up two or more responses. We have developed a set of rules which will take place during the data capture stage to resolve multi-tick responses.

### **Inconsistent Data**

5. The second type of error to be introduced is where data is inconsistent – the response to one question is not consistent with the response of another. We have developed a set of consistency checks. These represent our view of the world and of those instances we do not think can exist. There will however be some genuine cases but our assumption is that the majority of times these cases occur, they will occur in error rather than be true.

### **Invalid Data**

6. Invalid data may also be introduced where the data we receive is out of our recognised range, such as someone who says they are over 150 years old. These data will either be set to missing or corrected at the data capture stage.

### **Missing Data**

7. The final error which can occur in Census data is where people do not respond to a question when they are supposed to. All these data will be marked missing and will be dealt with using the donor imputation system described by Vickers and Yar (1998).

## **III. Resolving Inconsistent data**

8. There are four stages to resolving inconsistencies.

- i) Identify the inconsistencies
- ii) Resolve easy to edit records
- iii) Reduce the number of variables involved in inconsistencies
- iv) Resolve remaining inconsistencies using the donor imputation system.

### **III.1 Identifying Inconsistencies**

9. The first stage to resolving the inconsistencies is to identify those records that contain inconsistencies. This is done by comparing the household and person data against a set of consistency checks or edit rules. They can be split into two groups: those that occur within a person such as checks between age and marital status and those that occur between people in the household such as the age difference between parents and children. These rules essentially give our view of the world and consist mainly of explicit edit rules but also contain several implicit rules that will ensure that the donor system can find an appropriate donor if one exists.

10. In addition to these, the system will identify, but not resolve, records with combinations of variables that likely to occur in only a minority of cases, such as students who are aged over 55 years old. We do not want to correct these records but we do want to keep track of how often they occur for data quality purposes. These are known as soft consistency checks.

11. In all cases the consistency check that is broken and the variables involved in the conflict will be identified on each household record.

### III.2 Fixed edit rules

12. In 1991 inconsistencies were resolved using an edit matrix. This included a series of tables which identified inconsistencies in the data and either provided a valid value to be substituted for the invalid item or else marked an item for imputation. The tables were very large and were constructed to consider every possible combination of values and give the action required for that combination. This was a cumbersome process and was only performed on a subset of variables. It is not possible, nor desirable, to repeat this process for the 2001 Census. Firstly, we are planning to code 100% of the data and hence we would have to devise an even larger number actions and secondly we are intending to collect a large amount of relationship data between people within the household which makes the number of possible combinations escalate to an unwieldy size. More information on the 1991 Matrix can be found in Mills and Teague (1991).

13. However, there are some instances where it will be obvious what action to take and it would be sensible in these cases to invoke an appropriate edit rule. Therefore, we have developed some edit rules to be used in these cases. These have not been finalised. The outcome from these rules will be to either:

- i) change the value of one of the variables to a different value: or
- ii) set the value of one of the variables to missing.

### III.3 Reduction of inconsistent variables

14. The database will now consist of missing data and data with inconsistencies that have not been resolved by the simple edit rules. Before these are resolved by the donor imputation system the variables involved in the inconsistencies in each household are examined to identify any variables that are predominantly responsible for the inconsistencies. If a variable or variables are found then these are set to missing and will reduce the number of inconsistencies that are left.

15. For instance, suppose that variables A, B, C, D, E, F and G are marked as being involved in an inconsistency in a household. Further suppose that variable A is marked as inconsistent with B, C, D and E, variable B is inconsistent with D and E and F and G are inconsistent with each other. This is illustrated below.

**Table 1. Inconsistencies before reduction**

Inconsistency number	Variables involved in inconsistency						
	A	B	C	D	E	F	G
1	✓	✓					
2	✓		✓				
3	✓			✓			
4	✓				✓		
5		✓		✓			
6		✓			✓		
7						✓	✓
<b>Total number of times involved in an inconsistency</b>	<b>4</b>	<b>3</b>	<b>1</b>	<b>2</b>	<b>2</b>	<b>1</b>	<b>1</b>

Now given that A is involved in most inconsistencies it will be set to missing and the inconsistencies remaining will look like this.

**Table 2: inconsistencies remaining after A has been set to missing**

Inconsistency number	Variables involved in inconsistency						
	A	B	C	D	E	F	G
1	M						
2	M						
3	M						
4	M						
5		✓		✓			
6		✓			✓		
7						✓	✓
<b>Total number of times involved in an inconsistency</b>	<b>0</b>	<b>2</b>	<b>0</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>

16. In table 2 variable B is now involved in the more inconsistencies than any other variable and, therefore, will be set to missing as well. It is clear that once this happens only F and G will be involved in an inconsistency. Given that they will each be involved in the same number of inconsistencies (in this example 1 each) it will not be possible to identify which one is in error and so both will be marked as being involved in an inconsistency and set to missing before being processed by the donor imputation system.

#### III.4 Resolving inconsistencies using the Donor Imputation System

17. The resolution of errors within a recipient can be considered a four step process.

Step 1 consists of attempting to repair all errors within a household at once as described above

Step 2 goes on to simply correct one person at a time using the same donor approach.

Step 3 will correct any remaining between person inconsistencies

Step 4 will take place if it has not been possible to complete step 4 using a donor approach and will consist of a default method such as regression.

18. At the beginning of each stage the data will consist of clean records and records with missing data. Additionally these records can be broken down into three groups:

- Records consisting of missing data where the respondent did not provide an answer or missing data that has been created as a result of an edit rule or during the ‘reduction of inconsistent variables’ stage.
- Records where variables involved in within person inconsistencies have been effectively set to missing.
- Records where variables involved in between person inconsistencies have been set to missing.

19. The availability of a sufficient donor pool is essential for the system to work and so the system classifies the recipients into one of these three groups and then deals with each group in order. Once a recipient has been resolved it is added to the donor pool. Both the recipient and the donor have a weight attached to them so that they both have a lower chance of being used as a future donor.

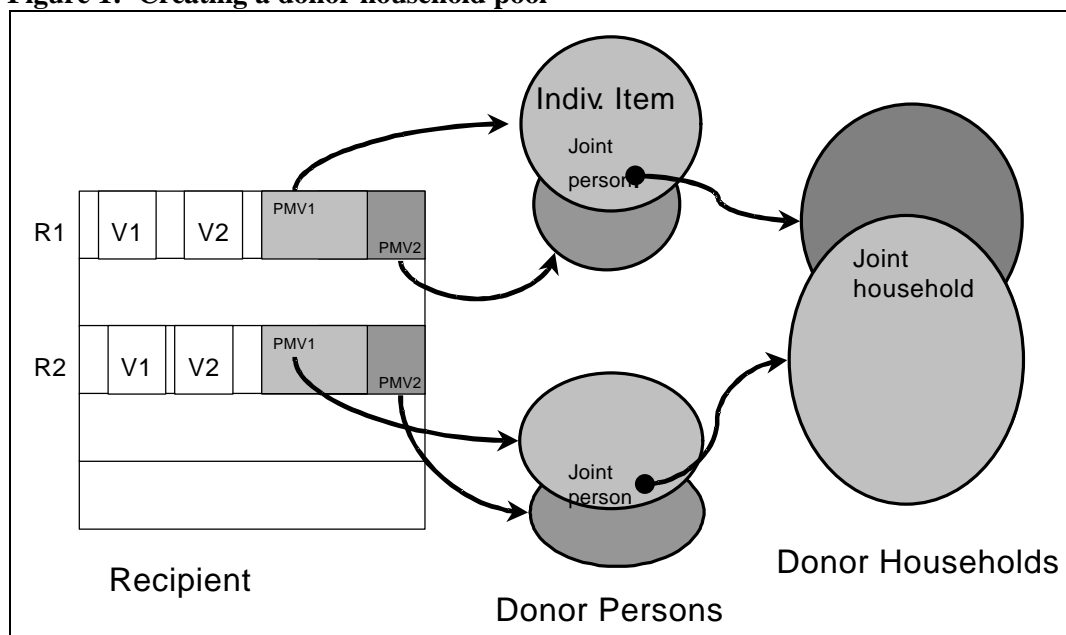
20. Searching for donors can be a very time consuming process and it is not necessary to consider, and therefore score, every record in the database as a potential donor. We, therefore, restrict the search to households of the same size. Additionally a record cannot be considered as a potential donor if:

- it fails any 'within-person' consistency check (those that fail soft checks can be considered);
- the variable(s) in error are missing in the potential donor;
- the PMVs are missing in the potential donor;
- it has already been used as a donor more than a certain number of times

21. The basic principal of the donor system is to search for and use a single donor for all the missing variables of a recipient record. The method for records with no variables involved in inconsistencies searches for a donor using a set of matching variables which are related to the missing variables of the recipient record. Primary matching which is used to identify a pool of potential donors takes place between all persons that have an error in the recipient household and the corresponding people in the donor household.

The following diagram (ONS (1999)) shows this process.

**Figure 1: Creating a donor household pool**



22. The left of the diagram shows that two persons in the recipient household have two missing variables. The PMVs for these people are used to identify potential donor persons that can resolve each missing variable. Those donor persons that can solve both missing values form the solution space for each person and those households that contain two different people that can resolve all missing variables form the solution space for the household.

23. This method is extended to treat records that have inconsistencies. For those variables involved in inconsistencies the system calculates a score to reflect the number of changes that are being made to these variables. If this score is above a minimum then the system will continue to search for a donor.

24. The system processes data in estimation areas which are the size of two or three local authority districts (500,000 people). The search for a donor is potentially a three stage process. However, the search will not proceed to a subsequent stage once a donor has been found.

- **Stage 1** searches for all potential donors within an estimation area which match the values of the matching variables in the recipient variable.
- **Stage 2** collapses the categories of the primary matching variables (PMVs) and searches the processing block for a donor. In collapsing the categories for each variable, we had to bear in mind the edit rules and did not merge categories involved explicitly in the edit rules otherwise consistency may not have been assured. In any case, however, we will apply consistency checks to confirm that no inconsistency exists after imputation.
- **Stage 3** reduces the number of matching variables.

25. If more than one donor is found during any stage for a recipient record, the donor with the least statistical distance based on secondary matching variables is chosen. Secondary matching, which is used to calculate the statistical distance between recipient and donor households takes place between variables for other persons in the recipient household and the corresponding variables for persons in the donor household. If, after this stage, there is still more than one donor, then the donor who is geographically closest is selected. This Donor System together with an evaluation of the system is explained in more detail in Vickers and Yar (1998).

26. At the end of stage three it is still possible that a donor household will not have been found to repair all the errors in the recipient household at once. If it has not been possible to find a donor household at step 1 the system will move onto step 2 and attempt to resolve persons in the household independently of each other. It will only continue to further steps if it has not been possible to correct the record at each of the preceding steps. Because the potential pool of donor households are put together by building up from each missing variable (figure 1), the information needed to identify a potential donor in steps 2 to 4 will already be available and these steps are likely to be less time consuming. We are still working on our default method and have not ruled out clerical resolution.

27. The system is currently under design but it is likely that scores for step 3 will be calculated before the scores for step 2, and likewise the scores for step 2 will be calculated before the scores for step 1. This means that it will be relatively straightforward to identify donors at each of the subsequent stages if a donor is not found during the preceding stage.

### III.5 Minimum Change

28. The system is designed to choose a donor that changes the minimum number of variables in a record. However, the donors available may not provide the minimum solution. This is unlikely to happen where the number of variables is small but may occur with a greater frequency where the number of variables involved is large. We will investigate next year how well the system obeys the principle of minimum change using data from the Census Rehearsal.

29. When calculating the changes that have been made to correct inconsistency we have assumed that a change of 5 years in age is equivalent to a change in categorical variable. For example, changing age by 5 years will be equivalent to changing marital status from single to married. This is the only weight that is applied to variables during the scoring process.

#### **IV. Conclusions**

30. This paper has given a description of the edit and imputation system which will be used in the 2001 UK Census. It shows that the system will be substantially different from that used in any previous Census and will place more emphasis on imputing and correcting data from other valid records within the dataset than ever before. There will be little or no manual intervention. The system will also endeavour to achieve minimum change although this is not guaranteed. By repairing the data using one donor it will ensure that the marginal and joint distributions within the data are preserved. The system has undergone substantial testing and this has been demonstrated in previous papers.

#### **Acknowledgements**

31. Many people have contributed to this system. I am grateful to Professor Ray Chambers, Southampton University for all his help and advice. Many colleagues at the Office for National Statistics have also provided invaluable help. These include Faith Anderson, Marie Cruddas, Andy Teague and Mohammed Yar.

#### **References**

Mills and Teague (1991). Editing and imputing data for the 1991 Census. *Population Trends*, Vol 64, pp 30-37.

Vickers and Yar (1998). The development and evaluation of the donor imputation system (DIS) for the 2001 UK Census of Population and Housing, Joint IASS/IAOS Conference, Mexico

Fellegi and Holt (1976). A systematic approach to edit and imputation. *JASA*, Vol 17, pp 17-35

ONS (1999). DEIS Study Report by Lockheed Martin