WP.17

ENGLISH
ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (iii): Editing and imputation of administrative and census data

EDIT AND IMPUTATION OF THE 2011 UK CENSUS

Supporting Paper

Submitted by the Office for National Statistics, UK[1]

I. INTRODUCTION

1. In 2001, the UK Census Offices processed about 27 million questionnaires containing almost 60 million people. Of these, 28 per cent of the person records contained one or more erroneous items which were corrected by the bespoke Edit and Donor Imputation System (EDIS). The 2001 system was based on the principle of minimum change proposed by Fellegi and Holt (1976). For 2011, the UK Census Offices have endorsed CANCEIS to perform robust, cost effective, editing and imputation whilst incorporating methodological best practice. Hence, CANCEIS forms the cornerstone of the Editing Strategy for the 2011 UK Census.

2. This paper outlines progress to date towards the implementation of CANCEIS in the 2009 Census Rehearsal as a stepping stone towards the Census proper. The general approach to methodological development has been to trial CANCEIS on samples of 2001 data, then once satisfactory results are achieved, the system is parameterised for 2009.

3. Section II of this document provides background information on the 2011 UK Census Editing Strategy, editing and imputation in the 2001 Census followed by an overview of CANCEIS. Section III describes the broad approach to methodological development of editing and imputation for the 2009 Census Rehearsal. Section IV describes trialling over-imputation as a method of statistical disclosure control. Section V describes plans for partial record level imputation as part of a coverage adjustment process and finally Section VI provides some concluding remarks.

II. BACKGROUND

2011 UK Census Editing Strategy

4. The 2011 UK Editing Strategy has the primary aim of imputing for all item level missingness and resolving inconsistencies in the responses for the households and persons affected (UKCDMAC(08)15). The Strategy is driven by three key principles:

[1] Prepared by Heather Wagstaff and Leone Wardman (Heather.Wagstaff@ons.gov.uk)

1. all changes that are made will maintain the quality of the data;
2. the number of changes to inconsistent data will be kept to a minimum; and
3. as far as possible, missing data should be imputed for all variables to provide a complete and consistent database.

5. The Principles and Strategy build on those used in the 2001 UK Census. The Principles are set within the wider perspective of the user requirements from the 2011 Census. By taking a strategic view we ensure that the planning and implementation of all methodological and operational work is driven by user requirements within an agreed quality framework.

Overview of 2001 UK Census Edit and Imputation

6. In the 2001Census, the UK Census Offices processed about 27 million questionnaires which related to 60 million people. Of these some 28 per cent of the person records contained one or more erroneous items. Over 13.7 million deterministic edits were carried out on the data for 11.8 million people. The eight most frequently executed deterministic edits accounted for 91% of the total. In addition 13.8 million people had one or more items imputed. The imputation process followed four main steps:

1. Joint (30%)	The Editing System searched for a complete household of the same size to act as a donor record. All missing or inconsistent values were repaired in the recipient by copying the values from the donor household.
2. Individual (67.38%)	If no donor household of the same size could be found, then each person in the household with edit failures was imputed individually.
3. Fallback (2.49%)	If the individual imputation failed then households up to size 8 were subjected to a semi-manual repair process where items were imputed one value at a time.
4 Clerical (0.13%)	If individual imputation failed and the households were over size 8 then they were repaired manually.

7. Overall, 30% of all households were processed using the joint imputation method; 67% by individual imputation; and the remaining 2.6% were repaired manually or semi-manually. For 2011, the UK Census Offices have endorsed the use of CANCEIS to perform robust, cost effective, edit and imputation whilst incorporating methodological best practice. There will be no deterministic edits as in 2001, but rather all erroneous values will be replaced stochastically, which will effectively double the amount of erroneous records to be repaired by the system. Hence, CANCEIS forms the cornerstone of the 2011 UK Census Editing Strategy.

Overview of CANCEIS

8. CANCEIS was developed specifically to perform editing and imputation for the 2001 Canadian Census and has been further enhanced for each subsequent Census. The system applies a joint imputation approach for the nearest neighbour imputation (NIM) of categorical and numeric variables. Its goal is to minimise the number of changes made to the recipient, given the available donors, while ensuring that the imputation actions are plausible according to a pre-specified set of user-defined edit rules. The edit rules are supplied in the form of Decision Logic Tables (DLTs) which are a highly efficient method of identifying inconsistencies and implausible values in the data. The joint imputation approach identifies donors for an entire household, not just for individual persons. Thus, CANCEIS implements a data driven approach: NIM searches for donors, and then determines the minimum number of variables to impute given the available donors. This is in contrast to the Fellegi-Holt approach where the minimum number of variables to impute is determined first, then imputation is performed by searching for donors.

Changing the order of these operations in NIM allows CANCEIS to solve larger and more complex edit and imputation problems (Bankier, 2000).

III. EDITING AND IMPUTATION OF THE 2009 CENSUS REHEARSAL

9. This section describes the broad approach to methodological development of editing and imputation for the 2009 Census Rehearsal. In 2001, 98% of household returns contained responses for between one and five people. This corresponded to the population who returned a single Household Form and are referred to as the main population. Hence, the development of editing methods has been separated into two parts: (i) methods for the main population; and (ii) methods for small populations which include large households, persons in Communal Establishments, and the Coverage and Quality Surveys. This section describes editing methods for the main population together with the broad approach to the imputation of persons in large households. Comment is also made about some of the underlying data issues identified from testing.

Micro-simulation approach

10. Initially methodological development was based on simulated micro-data which was formed from a sample of 2001 Census records. The simulation approach allowed a thorough analysis and evaluation of imputation performance. The 2001 sample was known as the reference data and contained household and person records for household sizes 1 to 9. Further samples of complete and consistent records were merged together to form a 'truth deck'. Next, the pattern of missing values present in the reference data was randomised 16 times within each household size and applied to the truth-deck. Hence, there were sixteen test-decks which were each imputed three times to account for imputation variance. Thus, each experiment contained 48 (16*3) replications. The Stuart-Maxwell statistic was calculated for each variable and for the bivariate distributions of key variables. A non-significant result indicated that the true and recovered distributions were not statistically different. Where a significant result was observed the results were closely scrutinised. Finally, the results of the 48 runs were combined to obtain a measure of how well the imputation procedure would be expected to perform under similar conditions. Initially, CANCEIS was parameterised to impute for the 9 household variables simultaneously; a similar approach was taken for the 23 person variables; and all CANCEIS parameters were fixed at their default settings.

11. The baseline results for the household variables were impressive. The Stuart-Maxwell test was non-significant for seven of the nine variables in over 72% of the experiments. The two variables returning significant results were highly correlated and closely scrutinised. For Accommodation Type, there was little predictive power to distinguish between the types of house (detached, semi-detached, terrace) or between the types of flat (purpose built, converted house, in a commercial property). The second, Self-contained Accommodation, is a highly skewed binary variable with an extremely low proportion of accommodation reported as not self-contained. The dominant category combined with unequal off diagonal elements in the transition matrix made the Stuart-Maxwell statistic unstable. Overall, the imputation process worked well for all household variables.

12. The baseline results for the person variables were rather mixed, which was unsurprising given the complexity of the question set. For 18 of the 24 variables, the Stuart-Maxwell statistic was significant in every experiment. Closer scrutiny of the results showed that this was often due to the instability of the statistic. However, overall, the quality of the imputations was poor. The results caused particular concern because the simulated data only contained missing values and had not yet been perturbed to include inconsistent responses. Thus, the proportion of erroneous responses was significantly lower than would be expected in a live environment.

Issues with the micro-simulation approach

13. The quality of the imputations for the household variables was surprisingly high and, given that a low level of inconsistencies existed in the reference data, was deemed sufficient to carry forward to the Rehearsal. However, distributional accuracy was not achieved for a high proportion of the person variables and the proportion of inconsistent responses was higher than expected. Over 55% of households contained missing values, inconsistencies or both; 36% contained missing values only; 8% contained inconsistent values only; and 10.5% contained both missing values and inconsistencies. The proportion of households which contained erroneous records increased with household size and ranged from 40% for household size 1 to 100% for household size 9. The patterns of inconsistencies observed in the reference data were extremely complex and resource constraints did not allow us to develop a method to perturb the simulated data.

14. All further work was based on a sample of raw 2001 Census data and household sizes one to five only. There were three factors which lead to the decision to limit the main method to five person households:

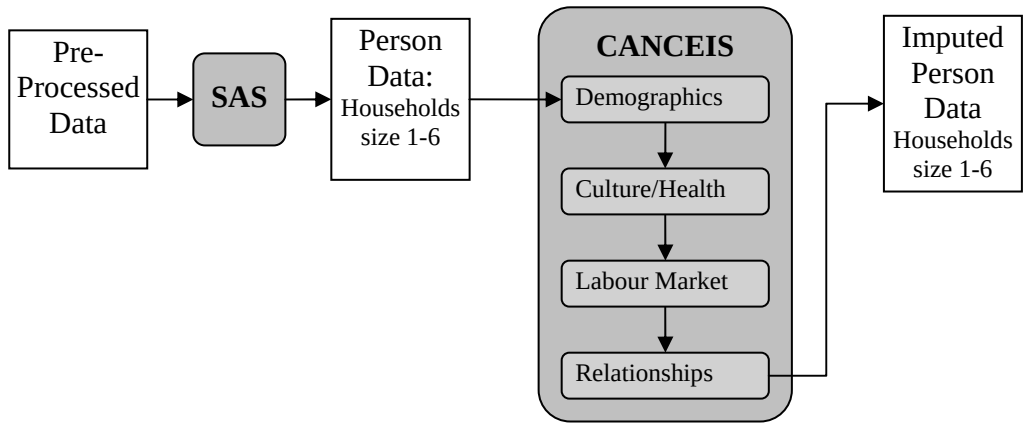
1. the low proportion of potential donor records (18% for household size 6 and 0% for size 9);
2. evidence of a clear break in data quality for six person households; and
3. in 2001 the main Household Form contained responses for one to five persons with additional persons recorded in Continuation Forms.

From this point forward we concentrated efforts on separating the 34 person variables into partitions according to their subject matter.

Partitioning by subject matter

15. Previous research had provided evidence that partitioning the person variables had the potential to significantly improve the quality of the imputed data (Wagstaff & Rogers, 2006). The person variables were separated into four partitions relating to both subject matter and the relative importance of each variable. The highest priority variables are those which relate to the population count and formed the first partition. In total, four partitions were formed which related to demographics, culture, labour market and relationship. The general approach to imputation of the four partitions is shown in Figure 1.

Figure 1: General approach to the imputation of Household Persons



16. In the 2009 Rehearsal, the data will be pre-processed: the household and person files will be sorted by stratum (household size), the data reformatted and the records for persons one to six will be written to the input file for the first partition 'demographics'. The output data from this run will become the input data in the second partition 'Culture/Health' but with the Demographics variables fixed as unimputable. This step is repeated for partitions three and four. The general approach of partitioning the person variables has been successfully implemented but with a number of underlying issues identified along the way. One of these related to the structure of the relationship matrix and another to the total number of edit rules required for each household size. Both of these issues are discussed below after which the broad approach to the imputation of households size seven and over is discussed.

Relationship matrix

17. The 2001 Census relationship matrix was specifically designed to meet United Nations and Eurostat recommendations to identify and classify family units within households. It also met UK Census users stated business needs to identify 'hidden' and 'reconstituted' families. There was a complete inter-relationship matrix for the first 5 people in a household; for the Continuation Form the matrix collected the relationship of a person to the two preceding people and to Person 1. For example it collected the relationship of Person 6 to Persons 1, 5 and 4. The 2009 relationship matrix is based on 2001, and contains inter-relationships for the first six people (since the main Household Form collects responses for up to 6 people). However, the Continuation Form collects relationships to person 1 and inter-relationships within the Continuation Form; for example Person 7 to Person 1, Person 8 to Person 1 and Person 7.

18. In order to show the complexity surrounding ensuring completeness and consistency of the relationship matrix, Table 1 shows the patterns of responses to the 2001 and 2011 questions. For example in 2001 and 2009, a household size 5 contains 10 directly observed relationships and 10 reciprocals. Similarly, a household size 10 contains: 25 observed relationships; 25 reciprocals; and 40 unspecified relationships which were constructed (20 of which are reciprocals). However, the patterns of the unspecified responses for larger households are substantially different between 2001 and 2009.

Table 1 2001 and 2009 Census Relationship Matrix

2001 Census

Person Number	Person Number									
	1	2	3	4	5	6	7	8	9	10
1	-	R	R	R	R	R	R	R	R	R
2	✓	-	R	R	R					
3	✓	✓	-	R	R					
4	✓	✓	✓	-	R	R				
5	✓	✓	✓	✓	-	R	R			
6	✓			✓	✓	-	R	R		
7	✓				✓	✓	-	R	R	
8	✓					✓	✓	-	R	R
9	✓						✓	✓	-	R
10	✓							✓	✓	-

2011 Census

Person Number	Person Number										
	1	2	3	4	5	6	7	8	9	10	11
1	-	R	R	R	R	R	R	R	R	R	R
2	✓	-	R	R	R	R					
3	✓	✓	-	R	R	R					
4	✓	✓	✓	-	R	R					
5	✓	✓	✓	✓	-	R					
6	✓	✓	✓	✓	✓	-					
7	✓						-	R	R	R	R
8	✓						✓	-	R	R	R
9	✓						✓	✓	-	R	R
10	✓						✓	✓	✓	-	R
11	✓						✓	✓	✓	✓	-

Key: ✓= collected; R = reciprocal; ☐ = derived

19. Consistency between responses to the relationship matrix is crucial for the derivation of the Household Composition Algorithm (HCA) which is a complex derived variable which seeks to identify and classify family units within households. Hence, the HCA must account for the multiplicity of modern day living arrangements and its accuracy is directly dependent on the

consistency of the relationship matrix. Overall, in 2001 it required excessive resource to deal with the complexity of editing, producing the HCA and quality assuring the output.

20. In the 2001 outputs, a number of relationships remained inconsistent with other variables due to two main reasons:

1. during imputation inconsistencies between relationship and other variables were usually resolved by changing the relationship matrix rather than other variables; and
2. in 6-person and larger households there was missing information and assumptions had to be made.

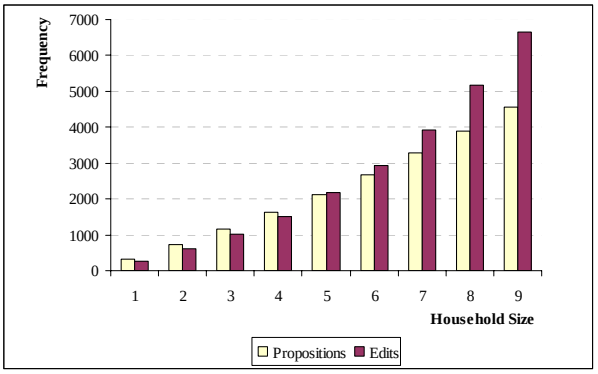
As shown in Table 1 above, there will still be missing information but the pattern will be different to 2001. Thus, for 2009 we are aiming to develop a fully automated process which is cost effective, efficient and negates manual processing. However, there are issues surrounding the number of edit rules which will be required.

Number of edit rules

21. The 2009 Rehearsal edit rules will be specified in CANCEIS as a series of Decision Logic Tables (DLT's). Within in each DLT the rows represent propositions and the columns are the edit rules. When developing methodology for 2011, ONS converted the 2001 edit rules into DLT's for household sizes 1 to 9. The number of DLT's propositions and edit rules are shown in Figure 1. which clearly shows that as household size increases so the volume of edit rules and propositions increases exponentially. For a household size 1, there are 36 DLT's which contain 256 edit rules and 323 propositions. Households size 9 need 74 DLT's containing 6,650 edit rules and 4,570 propositions. The cause of the steep increase is mainly from the volume of checks required to maintain consistency between the relationships.

Figure 1: Edit Rules, DLT's and Propositions
2001 England and Wales Census

Household Size	Number of Edit Rules	Number of DLT's	Number of Propositions
1	256	36	323
2	599	63	725
3	1,003	66	1,154
4	1,513	69	1,623
5	2,166	70	2,134
6	2,947	71	2,683
7	3,927	72	3,272
8	5,169	73	3,901
9	6,650	74	4,570
Total:	**24,230**	**594**	**20,385**



22. The Figure clearly shows that as household size increases so the volume of edit rules and propositions increases exponentially. For a household size 1, there are 36 DLT's which contain 256 edit rules and 323 propositions. Households size 9 need 74 DLT's containing 6,650 edit rules and 4,570 propositions. The cause of the steep increase is mainly from the volume of checks required to maintain consistency between the relationships. However, since the inter-relationship information for larger households is mainly contained within form type (except to person 1) the number of rules required is not as excessive as for 2001. However, a separate imputation method for the larger households still has to be developed.

Imputation of large households

23. Overall, in 2001 a significant amount of resource was required to manage the complexity of editing the relationship matrix and quality assuring the output. The majority of the effort was focussed towards larger households. In 2011 the ONS are seeking to eradicate the need for a manual process by taking a separate approach to the imputation of large households. The 2011 Household Form will collect responses for up to 6 persons. Assuming that 2001/2011 household distributions are broadly similar, we can expect about 99.5% of all households to return a single form. However, we acknowledge that, although a paper form will be posted to every household, internet data collection (IDC) adds an element of uncertainty to our estimates. Also, IDC is likely to further complicate the continuity problem, or break in data quality, between the Household and Continuation Forms. Larger households who choose to complete the paper form will have to contact the Census Helpline and ask to be forwarded one or more Continuation Forms, whereas IDC will automatically capture up to 30 people. It is crucial to analyse the Rehearsal data to provide an indication of the likely IDC take up rate and to ensure that there no indication of systematic error or bias present in one mode of collection that is not present in the other.

24. For the small proportion of households that return one or more Continuation Forms, ONS are taking a similar editing approach to that applied in the 2006 Canadian Census. Separate strata will be formed for households sizes 1 to 5 and then 6+ where the 6+ stratum contains 6 person households plus the first 6 persons of households size 7 and above. CANCEIS will be applied in three stages:

Stage 1: process persons 1 to 6 (where they exist) from all household sizes with the complete set of edit rules.

Stage 2: process persons 7 and over by removing the between person edit rules but ensuring that relationship to person 1 is imputed.

Stage 3: develop separate methodology to ensure consistency amongst the higher relationships.

The method is currently under development and will be thoroughly evaluated to ensure that it is robust and does not introduce an unacceptable level of distortion to the data.

IV. IMPUTATION AS A METHOD OF DISCLOSURE CONTROL

25. As part of the development of the overall 2011 UK Census programme, ONS has trialled over-imputation as a method of statistical disclosure control. Census outputs have a higher risk of disclosure and are harder to protect than other statistical data outputs. There are three main reasons for this:

1. they contain counts for the whole population;
2. small areas predominate in output geography; and
3. tables are disseminated from only one data source and it is straightforward to link and difference tables.

The key disclosure risk for the 2011 UK Census outputs is attribute disclosure which is defined as learning something new from the census data about an individual or group of individuals that was not previously known.

26. In readiness for 2011, ONS reviewed a wide range of statistical disclosure control methods of which three were short-listed for a thorough evaluation to assess risk and utility. The three methods were:

1. record swapping which involves perturbing the data by swapping the geographical identifiers of a small percentage of household records with other records by matching on specific control variables;
2. a cell perturbation method developed by the Australian Bureau of Statistics (ABS) which is a post-tabular method where table cell values have a perturbation added which is drawn from a 'look-up table'; and
3. over-imputation which involved introducing missing values into the data followed by their recovery, in this instance, using CANCEIS.

27. The research was conducted in two stages: the first compared geographic imputation, record swapping and the ABS method across three levels of perturbations (2%, 10% and 20%) with the two pre-tabular methods also broken down by whether perturbed records were selected at random or targeted. It should be noted that a 2% perturbation for swapping and over-imputation results in 2% of records being perturbed, whereas 2% perturbation for the ABS method results in 2% of cells in the table being perturbed. The second stage related to a non-geographic over-imputation which is described below. Finally, risk and utility measures were evaluated for a series of census output tables, each of which had different characteristics.

28. Two samples of 2001 Census data were identified each of which contained: population with differing characteristics; multiple Administrative Areas: and about 500,000 population. Each sample was stratified by Area and household size then a sample of households were selected at random. The strata ensured that over-imputation had some degree of comparability with record swapping since the latter was applied in two ways (1) swapping households between Administrative Areas; (2) by household size within Area. Thus all households had an equal probability of selection. The sampling scheme was repeated but the second selection targeted the population of high risk households.

29. Initially low level geography and age variables were blanked out and CANCEIS was applied using donors from the remaining population selected to replace the missing values. For the missing small geographies, donor records were selected from within the same Administrative Area whereas age was imputed using all possible donors within the sampled area. A second approach was to consider a non-geographic over-imputation method as it was felt that this was likely to provide better protection for both tables and microdata. As with geographic imputation, donors are found from the remaining population that match exactly or closely on other related variables. CANCEIS was applied to the data with all parameters fixed to the default setting and using only a minimal set of edit rules and basic distance measures.

30. The evaluation considered a number of factors including the disclosure risk, as measured by the proportion of small cells unchanged, which remained after the each of the methods had been applied. Briefly, it was found that for some tables, imputation perturbed more small cells than the ABS method but that record swapping appeared to protect more. Imputation had a larger effect on the variance than the other two methods especially with the targeted perturbations, which is as we would expect since high risk records are likely to be at the extremes of the distributions and have values imputed from donors which are closer to the mean for numerics or dominant values for coded responses. At the individual level, as we would expect over-imputation was found to have a greater impact on the relationships between variables where as record swapping did not (since the complete record is swapped between geographies). In terms of relative absolute differences, the ABS method had the lowest relative change whether random or targeted. On the whole, over-imputation performed slightly better record swapping for both random and targeted perturbations.

V. PARTIAL RECORD LEVEL IMPUTATION

31. Following the 2001 Census, records from the Census and Post Enumeration Survey (PES) were matched and formed the basis of a process to estimate the number of households and persons missed by the Census. The 2001 Census database was then adjusted to account for the estimated under enumeration. The adjustment process used the matched Census and PES dataset to derive coverage weights which were then calibrated to the estimates at the Administrative Area level. The weights were subsequently used to select donor households and people estimated to have been missed from counted households. Each imputed household was placed into either an empty household or into a random postcode within the Area; individuals who were missed from responding households were imputed into an existing household. A final phase adjusted the post-imputation Census database to ensure that the targets for household size and age-sex estimates were met exactly within each Area. The final process was problematic in some areas and often introduced spikes into the data for variables which were not directly controlled by the process.

32. In preparation for the 2011 Census, the ONS are trialling the use of CANCEIS for partial record level imputation. The coverage adjustment process will estimate the number and location of households and persons missed by the Census. That same process will insert skeletal records into the Census database at an Area level which contain the key variables used in the estimation process. CANCEIS will be applied in an attempt to impute for the missing items. Development and testing is scheduled for early 2010.

VI. CONCLUDING REMARKS

33. Over the coming months, ONS will translate the CANCEIS data dictionaries to the 2009 Rehearsal format together with the associated DLT's. ONS have also constructed a set of test data from 2001 records which have been reformatted to represent the 2009 variable structure. The Rehearsal questionnaire contains a number of new questions, for these their distributions have been simulated those observed in the social surveys. Once the live Rehearsal data is available the CANCEIS parameters will be optimised for use in the 2011 UK Census. However, there is still substantial amount of methodological development work to be completed to ensure the readiness of CANCEIS for the 2011 Census.

VII. REFERENCES

Bankier, M. (2000), "Imputing Numeric and Qualitative Variables Simultaneously", Social Survey Methods Division Report, Statistics Canada, Dated February 21, 2000.

Fellegi, I.P. and Holt, D. (1976) "A Systematic Approach to Automatic Edit and Imputation". Journal of the American Statistical Association, March 1976, Volume 71, No. 353, 17-35.

ONS (2003) "Census 2001: Quality Report for England and Wales" ONS.

Wagstaff, H.F. and Rogers, S.R. (2006), "Application of CANCEIS to 2001 Census data." Technical Report, ONS Titchfield, Internal Report.