

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE****CONFERENCE OF EUROPEAN STATISTICIANS****Work Session on Statistical Data Editing**

(Vienna, Austria, 21-23 April 2008)

**Topic (vi): Methodologies for the Editing of Census Data****TOWARDS THE 2011 UK CENSUS EDITING STRATEGY****Supporting Paper**

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

**I. INTRODUCTION**

1. The next Census of Population and Housing in the United Kingdom (UK) will take place in March 2001. It will be the 21st decennial Census to take place in the UK, the first was conducted in 1801. For the 1981 to 2001 UK Censuses, the Office for National Statistics (ONS) developed a series of bespoke automatic edit and imputation systems. Each of the systems was broadly based on the principle of minimum change proposed in the seminal paper by Felligi and Holt (Felligi-Holt, 1976). In 2004 ONS endorsed CANCEIS (CANadian Census Edit and Imputation System) as the corporate editing and imputation tool for sources where the data are mainly qualitative. CANCEIS provides cost-effective standardised edit and imputation whilst also incorporating methodological best practice (Bankier, 2000).

2. To date, ONS has successfully implemented CANCEIS on a number of household surveys and other statistical sources including data from the registration of Life Events. Based on the outcome of an applied research programme, CANCEIS has been endorsed for the 2011 UK Census (Wagstaff et al, 2004; Wagstaff and Rogers, 2005). Hence, CANCEIS forms the cornerstone of an extensive research programme which aims to deliver a robust and coherent 2011 Census Editing Strategy. Initial findings indicate that the research will lead to a robust strategy that is efficient and has the potential to produce complete and consistent data of a very high quality.

3. This paper presents an overview of the approach being taken to develop the 2011 UK Census Editing Strategy. To set the scene, Section II provides a brief overview of the editing process in previous UK Censuses; Section III provides an overview of CANCEIS; Section IV outlines the research which led to the endorsement of CANCEIS for the 2011 UK Census; Section V describes the development of the 2011 Editing Strategy and the associated research programme. Finally, Section VI provides concluding remarks.

---

<sup>1</sup> Prepared by Heather Wagstaff and Steven Rogers ([Heather.Wagstaff@ons.gov.uk](mailto:Heather.Wagstaff@ons.gov.uk))

## II. OVERVIEW OF EDITING IN THE UK CENSUS

4. The Census of Population and Housing is conducted every 10 years in the UK. From the 1961 UK Census onwards, the adoption of electronic data processing techniques provided scope for the thorough and systematic checking of census returns. For the 1981, 1991 and 2001 UK Censuses, ONS developed a series of hard coded automatic editing and imputation systems each one broadly based on the principle of minimum change (Felligi-Holt, 1976). In 1981 and 1991 a sequential imputation method was applied which imputed the missing values individually and in a predetermined order. The 1991 system ensured that within-household consistency was maintained through a matching process which was based only on the variables included in the edit rules. However, evaluation of the 1991 system identified two key concerns: firstly, the joint distributions were not always maintained; and secondly, sequential processing did not make full use of all the available data. Further, the final 1991 Census dataset contained item level non-response which were coded as 'not specified'.

5. Prior to the 2001 Census, a lengthy consultation exercise took place which formed the basis of a careful assessment of information requirements. Discussions with Users of census data, clearly emphasised their requirement for complete and consistent outputs. As a consequence, the 2001 Census Editing and Imputation Strategy was developed with the primary aim of imputing for all missing data and resolving inconsistencies in the responses for the households and persons affected. The Strategy followed four basic principles:

1. all changes that were made would improve the quality of the data;
2. the number of changes to inconsistent data would be kept to a minimum;
3. as far as possible, missing data would be imputed for all variables, to provide a complete and consistent database;
4. the system had to be relatively easy to develop and capable of processing large amounts of data automatically within short timescales.

6. To achieve the aims of the Strategy ONS developed the bespoke Edit and Donor Imputation System (EDIS). The system requirements were specified in two parts:

1. the edit process which scrutinised every household and person record to identify inconsistent fields; and,
2. the imputation process which sought to implement a joint imputation method.

Following the imputation process all records were passed through the edit rules a second time to ensure that consistency had been maintained. Although EDIS was designed for the 2001 UK Census, each of the four countries (England, Northern Ireland, Scotland and Wales) had slightly differing requirements. Since EDIS was hard coded, the consequent variations in form design and editing requirements lead to a significant amount of resource and effort being spent on ensuring that the processing for each country was fully specified and implemented to the agreed quality standards.

7. During live running, EDIS performed fully to specification and within planned timescales. Imputation accounted for the bulk of the running time. A total of 13.7 million edits were carried out on the data for 11.8 million people. The base population for EDIS was 49.4 million people in England and Wales and the eight most frequently executed edits accounted for 91% of the total. One or more items needed to be imputed for 28.0% (13.8 million) of the population who returned Census forms.

8. However, during the 2001 processing operation, problems arose from three key sources: firstly, complex filter questions were not followed by a number of respondents; secondly, late changes to the question set went into the live operation untested; and, thirdly, the lateness of captured data from the 1999 Census Rehearsal gave no opportunity for system testing. The lateness of the Rehearsal data precluded testing EDIS on real census type returns and the opportunity to identify problems with form completion was lost. As a consequence, ad-hoc changes to EDIS were made during live running (ONS, 2003). For the 2011 Census, rather than design and build a further bespoke system, the UK Census Offices have taken the innovative step of endorsing CANCEIS.

### **III. BACKGROUND TO CANCEIS**

9. The ONS Statistical Modernisation Programme (SMP) focussed on the delivery of a standard technical infrastructure, methodologies and statistical tools. The main aim of the SMP was to identify and apply common recognised standards and practices in a highly efficient way. CANCEIS was endorsed as one such tool with the objective of implementing cost-effective standardised editing and imputation, whilst also incorporating methodological best practice.

10. CANCEIS was developed specifically to perform editing and imputation for the 2001 Canadian Census. The system allows data from a single donor to be used 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. CANCEIS implements a joint imputation approach, that is to identify 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).

11. To date, ONS have implemented CANCEIS on several household surveys and other statistical sources including data from the registration of Life Events. However, the associated datasets are all relatively small when compared to the volumes of data expected in a live census operation. Consequently, ONS sought assurance of its robustness to support the 2011 UK Census.

### **IV. ENSURING THE FITNESS OF CANCEIS FOR THE 2011 CENSUS**

12. The 2001 UK Census operation processed about 27 million forms containing information in respect of almost 60 million people. Hence, even though ONS had endorsed CANCEIS as the corporate edit and imputation system, it was crucial to provide assurance of its capability to support the 2011 UK Census. To provide such assurance, applied research was undertaken which aimed to assess whether CANCEIS:

1. could produce clean and consistent census data (provide proof of concept);
2. contained the functionality to fully replicate the 2001 Census Editing Strategy; and
3. contained an imputation process of acceptable quality, as measured by recovery of the statistical properties of the 2001 Census data.

#### **Providing proof of concept**

13. This part of the research aimed to provide proof of concept that CANCEIS contained sufficient functionality to produce clean and consistent census data. This was demonstrated by a somewhat simplistic approach: the 2001 edit rules were identified and re-written as DLT's; we then applied CANCEIS to raw 2011 Census data containing just over 171,000 households and 410,000 population. The results provided evidence that, given a set of correctly specified DLT's, CANCEIS easily identified and corrected for all inconsistencies and missing values as defined by the set of edit rules. The validity of the editing process was confirmed by comparing the outputs from CANCEIS against the outputs from coding the edit rules independently in SAS (Wagstaff, et al, 2004). However, although the CANCEIS outputs were complete and consistent, it was not possible to ascertain, with certainty, whether the statistical properties of the data had been maintained.

### Replicating the 2001 Census Editing Strategy

14. The broad range of functionality available in CANCEIS was demonstrated by replicating the 2001 Editing Strategy in a micro-simulation environment. The 2001 Editing Strategy was comprised of some 13 steps: 5 sets of deterministic edits followed by some derivations, 3 sets of consistency checks followed by derivations, and a 4 staged imputation process. The deterministic edits included a series of manipulations and validation which were executed before EDIS could be run. Table 1 shows the steps of the 2001 Strategy.

**Table 1. Replicating the 2001 Census Edit Strategy in CANCEIS**

2001 Census Edit Strategy	Deterministic Edits	Edit and Donor Imputation System	CANCEIS
1 Multi-tick rules	✓		✓
2 Range checks	✓		✓
3 Reciprocal relationships	✓		✓
4 Filter rules	✓		✓
5 Derived Variables	✓		✓
6 Within person checks		✓	✓
7 Between person checks		✓	✓
8 Soft consistency checks		✓	✓
9 Derived variables		✓	✓
10 Joint imputation		✓	✓
11 Individual person report		✓	✓
12 Fall back		✓	✓
13 Manual imputation		✓	✓

15. It was straightforward to implement all of the 2001 Strategy in CANCEIS thereby clearly demonstrating the broad range of available functionality. Thus the deterministic edits could be easily applied in the CANCEIS derive engine making efficiency gains by achieving all steps in a single system. However, although it is straightforward for CANCEIS to complete all the required steps in a research environment, it may not be operationally efficient to do so in a live operation the size of the UK Census (Wagstaff and Rogers, 2006).

### Recovery of the statistical properties of the data

16. The final stage of the research assessed whether the CANCEIS imputation process was of sufficient quality to be implemented in the 2011 UK Census. This was achieved by evaluating how well CANCEIS recovered the statistical properties in a micro-simulation environment. The micro-data were constructed by analysing a set of 2001 Census reference data, then replicating the observations using other complete and consistent census records. The resulting ‘synthetic’ data contained hierarchical records in respect of about 170,000 households and 400,000 individuals which formed a ‘truth deck’ with known statistical properties. Finally, we perturbed the synthetic data by introducing missing values which reflected the patterns of bivariate missingness observed in the reference data.

17. Since CANCEIS implements a stochastic process, we expect to observe a small element of variation in the outputs when repeating the imputation under the same conditions. To account for the variation, CANCEIS was applied to the micro-data 30 times. The quality of the recovery of the properties was measured by the Stuart-Maxwell test statistic for distributional accuracy and the Kappa coefficient, in combination with proportions of true values recovered, for predictive accuracy. The results of the imputation runs were then scrutinised to decide whether the outputs were acceptable.

18. Overall, the analyses provided strong evidence that CANCEIS recovered the marginal and joint distributions extremely well. Even at the lower bound of performance, as measured by the lowest recovery of the distributions, there was no evidence of significant differences between the ‘true’ and imputed data. The poorer recoveries did not arise from CANCEIS, but rather from two differing sources: records with rare characteristics with few or no matching donor records; and where the matching variables did not discriminate well. For example, there was little information amongst the predictors to discriminate between males and females. Where this occurred, as expected, CANCEIS apportioned donor records according to the proportions observed in the donor pool. Following the

successful completion of the research programme, CANCEIS was formally endorsed by the UK Census Offices to be applied in the 2011 UK Census.

## V. 2011 CENSUS EDITING STRATEGY

19. For the 2011 Census, a number of major changes in data collection and processing methods are planned. For the first time, census forms will be mailed out to a large proportion of the population; and, all households in the UK will be offered the option to complete their forms via a secure internet based application. Both of these changes pose significant challenges for the editing process. By adopting CANCEIS for the 2011 Census, ONS has the opportunity to research the potential impact of these changes and a number of other methodological issues.

20. The structure and content of the 2011 Census paper questionnaire will soon be finalised. The overall structure of the questionnaire is similar to 2001, a 'pages per person' format containing three sections: questions about the household; a relationship matrix; followed by a set of person questions which include skip patterns. The ONS aim to design the Internet questionnaire in such a way as to minimise response bias. Thus, the paper and internet questionnaires will be identical in terms of question wording, instructions and response categories. However, the format of the Internet questionnaire is likely to differ from the paper version and this difference may affect how people respond to the questions. Based on the experiences of other NSI's, we expect to observe some, as yet unknown, degree of bias between the two collection modes.

21. An applied research programme is now underway which aims to deliver a coherent editing strategy for the 2011 UK Census. Since the 2011 form has not been formally endorsed, the edit and imputation strategy is being developed in a staged approach. A series of methodological topics have been identified to be researched over the next two years. As the results of each topic becomes known we will seek formal endorsement, through the Census Governance Structure, before updating the Strategy. The lessons learned from EDIS, applied in 2001, were well documented and have been taken forward and addressed in the 2011 Editing Strategy. For example, the variations in questions between the 4 UK countries can be easily managed by CANCEIS as it is parameter driven rather than hard coded. A selection of the methodological issues from the research programme are described here.

### **Development of Imputation Methods for the Main Population (Persons and Households)**

22. Partitioning Variables: The 2001 question set contained 9 questions about the household, a relationship matrix and 34 person questions. The 2011 question set is likely to be similar number but will also contain a number of new and more complex questions. Traditional imputation methodology, based on the Fellegi-Holt principle of minimum change, tend towards the use of a single donor for the imputation of all missing values in a recipient. The principle is supported by CANCEIS which implements a joint imputation method, identifying donors for an entire household, not just for individual persons. However, there is concern that the single donor approach may not be the optimal strategy given the size and complexity of the 2011 Census question set. Matching variables can be identified and implanted that are specifically related to each topic thus improving the accuracy of donor selection. Furthermore, as potential donors are typically chosen from records with no missing values, partitioning the full variable set can also serve to increase the size of the donor pool. Hence, research is underway to establish the feasibility of imputing the person questions in groups relating to discrete topic areas (demographics, ethnicity & country of birth, health & wellbeing, labour market etc). The results are evaluated against a baseline constructed from the traditional single donor approach. Any observed trade-off will be quantified and reported.

23. Permutation of Household Persons: From evaluating responses to the 2001 Census, we observed that the ordering of persons within households is not necessarily consistent. We are researching whether the quality of donor records, and speed of processing, can be optimised by re-ordering persons within households. To do this we need to ensure that the 'household reference person' is placed in the first position with subsequent persons re-ordered in a structured way. We aim to determine whether it is feasible to separate families within households before reordering persons within the families to maximise the quality of potential donors.

24. Migration: The UK Census asks for information about the respondents address one year before census night. The question asks for the country of usual residence for international migrants and post-code for internal migrants. This question has proved problematic in the previous census. In the UK, postcodes and associated addresses are held on a database which is compiled by the Royal Mail. To date our research has focussed on statistical closeness whilst geographical closeness adds a further dimension.

### **Ensuring Robustness of the imputation process for Small Populations**

25. Large Household Sizes: As household size increases above size 2 so the number of households at each size decreases monotonically. This in turn implies a potential for sparseness of donor records at the larger household sizes. In consequence, a joint imputation approach becomes less plausible. To date have explored merging together households of sizes 6 and 7 to provide a larger dataset and proof of concept of the approach. It was necessary to append a dummy variable to represent the 'missing' person record in the households size 6 to maintain a rectangular matrix for CANCEIS. As an alternate approach, we will consider imputing at the person level for extreme donor sparseness. However, this approach leads to challenges when maintaining between-person consistency.

26. Communal Establishments: In the UK Census, persons resident in Communal Establishment's (CEs), or Collectives, have historically been treated as single person households. Where possible, donors have been selected from amongst the other residents within the same CE before seeking individuals from a different CE of the same type. The overall process of non-related persons will be researched to understand whether it is preferable to seek donors from a different CE, of same type and approximate size, as that for the recipient individual.

### **Essential Implementation Work**

27. Differing area types: The existing micro-simulation environment was developed using 2001 Census data drawn from Administrative Areas identified to be broadly representative of the UK population. There is a need to understand the volumes and patterns of missingness in other area types, for example inner cities, inner London and rural areas. Decisions can then be made about whether the existing synthetic data is adequate to continue the research or whether further synthetic datasets should be constructed for further research.

28. Imputation Diagnostics: In 2001, one or more items needed to be imputed for 28.0% (13.8 million) of the population who returned Census forms. Of these, 4.7million (m) were dealt with by joint imputation and 10.0m were imputed using individual imputation, including all those in single person households. For household variables, 2.5m needed imputation, 11% of all households. 0.08m were dealt with by fallback and the remainder by joint imputation. Almost all the donor households for joint imputation were used once each. These are examples of why the imputation process requires a set of diagnostics which provide information about the quality of the imputations. Under this heading is also the meta-data provided to Users of the census data. The starting point for both these pieces of work is the 2001 processes which will be reviewed. The diagnostics will also identify non-imputable records. There will be many instances where records are of such poor quality that it is not possible to impute them automatically. When this situation arose in 2001 it was dealt with by a procedure known as Fallback. The specification for the Fallback process will be revisited as a basis for this work and then options for 2011 researched.

## **VI. CONCLUDING REMARKS**

29. A programme of applied research has clearly demonstrated that CANCEIS contains an extensive range of functionality and is sufficiently robust to support the edit and imputation of the 2011 UK Census. The adoption of CANCEIS for 2011 has allowed ONS to focus on solving complex methodological issues that might have otherwise been neglected. CANCEIS is a highly flexible tool since it is parameter driven and operates on a series of user defined decision logic tables, as opposed to being hard coded. During the 2001 Census processing operation, problems arose from four sources: differing requirements for the four UK countries; complex filter questions not understood by some respondents; late changes to the question set; and the lateness of captured data from the 1999 Census

Rehearsal. Whilst we might plan to ensure that these same issues do not occur in 2011, the inherent flexibility of CANCEIS serves to mitigate the associated risk to the consistency of the data.

30. An on-going applied research programme aims to deliver a robust and coherent edit and imputation strategy for the 2011 UK Census. Since the 2011 form is not yet specified, we are developing the edit and imputation strategy in a staged approach. A series of methodological topics have been identified as requiring research over the next two years. As the results of each topic becomes known we will seek formal endorsement, through the Census Governance Structure, before updating the Strategy. Initial findings indicate that the research programme will deliver a Editing Strategy that is reliable, efficient and yield corrected data of an exceedingly high quality. However, there is still much work to do and CANCEIS is simply the tool by which to achieve the aims.

## **VII. REFERENCES**

Bankier, M. (2000) "2001 Canadian Census Minimum Change Donor Imputation Methodology". Working Paper No. 17, UN/ECE Work Session on Statistical Data Editing, Cardiff.

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) "Evaluating the 2001 UK Census". Working Paper No.11, Joint ECE-EUROSTAT Work Session on Population and Housing Censuses (Ohrid, The former Yugoslav Republic of Macedonia, 21-23 May 2003).

Wagstaff, H.F. and Skentlebery, R., (2004) "Report on the initial evaluation of CANCEIS on 2001 Census data". ONS Internal Report.

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

Wagstaff, H.F. and Rogers, S., (2006) "Diagnostics for the evaluation of imputed data". Working Paper No.6, UN/ECE Work Session on Statistical Data Editing, Bonn.

Wagstaff, H.F. and Rogers, S., (2007) "Optimising the 2011 UK Editing Strategy, not Re-inventing the Wheel". ISI Southampton, Special Topic Session on Census.