WP. 43
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (vii): Editing and imputation of census data

The Practical Implementation of the 2011 UK Census Imputation Methodology

Prepared by Stephanie Aldrich, Leone Wardman and Steven Rogers, Office for National Statistics, United Kingdom

I. Introduction

1. The 2011 Census was the largest statistical operation ever undertaken in the UK achieving response levels at least as good as in 2001. The first release of results was recently delivered on time and on budget. Within the wider census operations, the edit and imputation strategy aimed to develop an appropriate statistical methodology for imputing census data which not only produced a complete and consistent database, but also satisfied the technical aim of integrating into a wider automated system.

2. This paper aims to outline the development of the 2011 edit and imputation methodology and how the system performed in practice. Section II gives some background into the development work from the 2001 Census to the final 2011 system, while Section III gives an overview of the tuning process and the changes made in response to the live data. Section IV explains the fallback methods required to produce a fully imputed database and finally, a preliminary evaluation of the 2011 edit and imputation system is given in Section V. A sister paper (Wardman, Aldrich & Rogers, 2012) describes how the edit and imputation system was integrated into the automated census operations including the diagnostic process involved in evaluating the results.

II. The 2011 Edit and Imputation Strategy

3. The overarching aim of the 2011 edit and imputation strategy was to impute all item level missingness and resolve all inconsistencies while minimising changes to observed data and maintaining data quality. There was also an underlying aim of improving on the results from the 2001 edit and imputation strategy. In 2001 a bespoke Edit and Donor Imputation System (EDIS) was created which first ran a series of deterministic edits on the data and then imputed the data to obtain a complete and consistent database. The system first attempted a joint donor imputation method with a single donor household for each failed household. This is desirable because it maintains joint distributions, which is important for future analysis and allows for the application of between person consistency editing. However in 2001 only 34% of the population was imputed using this method. Alternative methods included individual person imputation, the replacement of failed households with clean households and manual imputation. With over one million people having been imputed using more than one method there was also a lack of consistency in the 2001 methodology. In addition, each data processing unit of around half a million people took 48 hours to be processed by EDIS. With more data expected to be processed in 2011, the 2011 edit and imputation system aimed to be faster than EDIS as well as delivering a higher rate of joint imputation while maintaining consistency in the methodology.

4. Following the 2001 Census, early research by Wagstaff & Rogers (2006) identified CANCEIS (Canadian Census Edit and Imputation System) as an appropriate tool with the potential to fulfil the methodological aims of the 2011 edit and imputation strategy. This package applies a nearest neighbour donor imputation method specifically designed for census data which handles categorical, numeric and alphanumeric data simultaneously. The software is specifically designed to conduct joint household imputation while editing and imputing simultaneously, maintaining between person consistency through the application of edit constraints. Testing based on a synthetic set of 2001 Census data showed that CANCEIS was excellent at estimating the unobserved distributions and flexible enough to ensure a consistent method was applied to data from the different countries in the UK and the different questionnaire types used in 2011. During testing, CANCEIS also showed improved processing speed by achieving in several hours what EDIS took two days to achieve. The other advantage of CANCEIS is the ability to execute it from various platforms, such as SAS or JAVA. This meant it could be easily embedded into the automated census Downstream Processing (DSP) system; a bespoke product built in house specifically to process the census data, which was one of the technical requirements for 2011.

A. Development of the 2011 Census Edit and Imputation Method

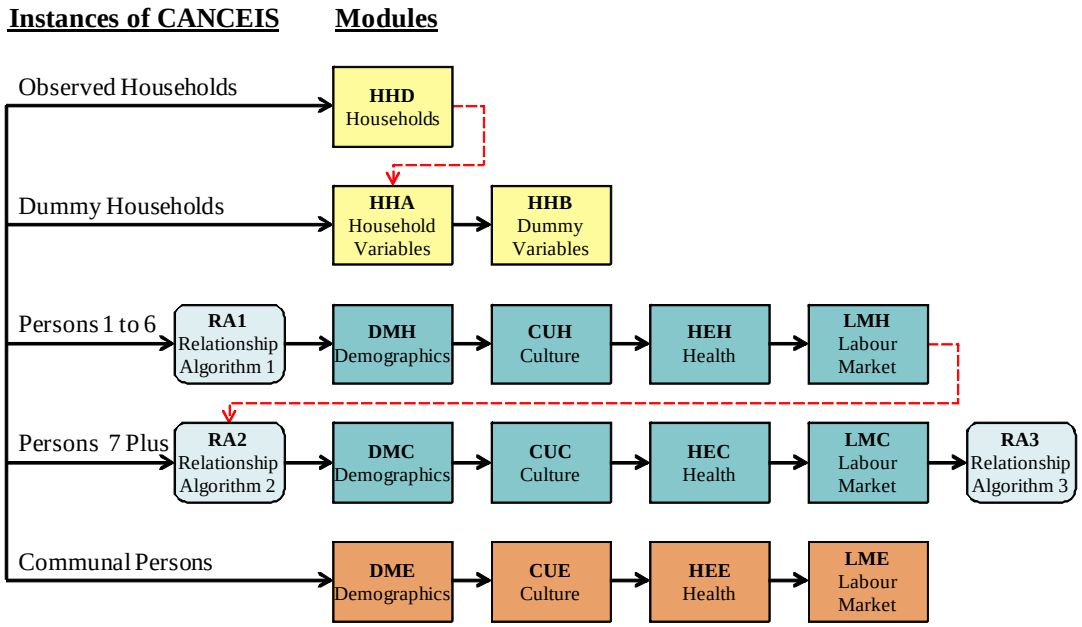
5. The final 2011 imputation method took several years to develop, initially using 2001 data and later using a modified subset of 2001 census data which was augmented with social survey data to approximate the data expected from the 2011 Census. While several issues were uncovered and resolved during the development; further problems arose during live operations that could not be anticipated because they were specific to the observed 2011 data. The following sections outline the development of the method and describes the main problems encountered both during development and later in live operations, along with the solutions applied to overcome them.

A.1 Breaking the data into imputation groups

6. With a responding base population of 53.5 million people to process, it was not practical to impute the whole database in one pass; firstly because the data was delivered to ONS by geographical region, or processing unit (PU). In addition, the donor pool would have been reduced below a viable level as just one missing value in the 43 person level questions would remove the whole household from the pool, compromising data quality. Therefore the data was processed through DSP in 101 geographical PUs containing on average 241,000 households with 530,000 person records. To ensure a timely delivery of results the PUs were processed by edit and imputation in parallel. Each PU was divided into the 15 imputation groups depicted in Figure 1. The five rows represent a run, or instance, of CANCEIS and equate to the data from each questionnaire type. For example, questions on the household accommodation were imputed separately from the questions about the individuals living in it. Finally, within each instance of CANCEIS, the data was imputed in a series of linked modules, which increased the donor pool for a given set of related variables.

7. The first two instances of CANCEIS in Figure 1 refer to the imputation of household data. Information was collected for two types of households; observed households that returned their questionnaire, and dummy households which refer to ‘place holder’ responses completed by field staff where no questionnaire was returned. There were only nine questions for observed households which could all be included in one module of CANCEIS, the HHD module. The dummy household variables, which included a subset of observed household questions that the enumerator could determine by sight or through a neighbour and dummy specific variables, had to be imputed in two modules of CANCEIS. After the observed households had been imputed, the records that required no imputation were used as donors for the first dummy household module (HHA) to impute the household questions while the second dummy household module (HHB) contained the dummy specific questions that could only be imputed using other dummy records as donors.

Figure 1: 2011 Census Editing and Imputation Process



8. The next three instances depict the person level imputation where the 43 person questions asked to every individual in a household or a communal establishment were imputed. The person instances of CANCEIS were dictated by the type of questionnaire the person answered; a communal person questionnaire, the main household questionnaire (containing persons 1 to 6) or a household continuation questionnaire (containing persons 7 to 30). In 2001 it was observed that the between person coherence dropped significantly for the persons answering on a continuation questionnaire so it was decided to only impute persons on the main household questionnaire as a joint imputation; that is, with a single donor household of the same size. Those on a continuation questionnaire were imputed as individuals because of the difficulty in achieving a joint imputation for larger households. To ensure consistency between the persons in the two instances, key variables for person one, such as age, sex, marital status and their relationships to persons two to six, were included in the person 7 plus imputation. Households that contained more than 30 people were converted to communal establishments and imputed as individuals with the rest of the communal responses in the final instance of CANCEIS.

9. With the addition of several derived variables used for matching, there were up to 58 variables for each person level instance to be imputed. This was too many variables to be imputed at the same time and would have reduced the donor pool significantly so a modular approach was adopted. Each person level instance of CANCEIS was processed in four imputation modules; demographics (DM), culture (CU), health (HE) and labour market (LM). The imputation modules were governed by both the implicit and explicit questionnaire routing as well as inter-variable relatedness, while the order of the modules was dictated partly by the questionnaire filters and partly by the logic of imputing the more accurately captured demographic variables upfront. Initially, logistic regression was used to find a set of five key variables which were good predictors of each other; age, sex, marital status, relationship to person one and economic activity. These were all imputed in the first module then fixed and used to help find donors in the later modules. However, due to the questionnaire filters a number of variables on second residences were also included in this module. (A more detailed discussion of the modular methodology is in a sister paper Wardman, Aldrich & Rogers, 2012). There were very few difficulties during the household level imputation, however the person level imputation proved more problematic. The following sections describe some of the issues encountered in the development and implementation of the person data.

A.2 Relationship Matrix Imputation – Additional Editing

10. Since CANCEIS edits and imputes simultaneously, the deterministic edit process of 2001 was not required in 2011. The exception to this was the relationship data, which is the most complex question in the Census. Family units within households were identified by responses to the relationship questions.

Each household member was required to record their relationship to the household members listed on the questionnaire before them using person one as the household reference person for continuation questionnaires. Figure 2 shows the pattern of observed relationships, for example, person seven should have an observed value to person one but not for persons two to six. The relationships imputed by CANCEIS are indicated by a cross.

11. During development it was discovered that observed values in key variables would often be changed instead of relationships when imputing errors in the relationship data. Therefore it was vital to stabilise the relationship matrix before processing the data by CANCEIS. To resolve some of the more common relationship errors, a relationship algorithm (RA1 in Figure 1) was developed in SAS that applied a series of deterministic edits, for example correcting for where a parent and child relationship had been recorded the wrong way around. RA1 ran before the person 1 to 6 imputation and only changed a value under strict criteria where there was a high level of certainty that an error was recorded. A second algorithm (RA2) ran between persons 1 to 6 and persons 7 plus imputation to amend the higher order relationships with person one's imputed relationships. A final relationship algorithm (RA3) ran after the persons 7 plus imputation to impute the remaining invalid or inconsistent values in the higher order relationships which were not imputed by CANCEIS (indicated by a circle in Figure 2).

Figure 2: Subset of observed relationships

Person \ Relationship to person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	···
1																	
2	x																
3	x	x															
4	x	x	x														
5	x	x	x	x													
6	x	x	x	x	x												
7	x																
8	x						o										
9	x						o	o									
10	x						o	o	o								
11	x						o	o	o	o							
12	x																
13	x											o					
14	x											o	o				
15	x											o	o	o			
16	x											o	o	o	o		
⋮																	

A.3 Records Remaining Unimputed - Fallback Imputation

12. During development the CANCEIS parameters were tuned to the modified 2001 test data and it was found that several records remained unimputed from each module. Although it was anticipated that tuning to the 2011 data would minimise the number of unimputed records in the automated system, a fallback method of imputation was necessary to meet the requirement of a complete and consistent database. A process was built into the automated system which allowed for records to be imputed outside of DSP and updated manually in the database. The intention was to apply the same method that was used in DSP (shown in Figure 1) but to adjust the CANCEIS donor search parameters to increase the chances of finding a donor that would satisfy all the edit constraints. The following section outlines the tuning process while the fallback imputation method is discussed in Section IV.

III. Tuning the System in Live Running

13. Around six weeks after census night the first area was processed through the edit and imputation system. Despite the successes in development, it took five months to stabilise the database, tune CANCEIS and successfully impute the first PU in live operation. This was partly due to the system being heavily reliant on the data arriving in the expected format from the earlier processes and predominantly

due to the live data being quite different from the modified 2001 test data. During the tuning period any data discrepancies from the earlier processes were resolved and the CANCEIS imputation parameters were adjusted in response to the live data by processing several PUs through the system from different areas of the country. An inner London PU, a wide spread city PU and a rural PU were used in order to get a representational cross section of the country. Several elements of the tuning process are outlined in the following sections.

A. Data Irregularities

14. Edit and imputation was the first statistical process the data reached in DSP and was the first point where between variable inconsistencies were identified. Although basic value checking was carried out during previous processes, for example checking that values were within their prescribed ranges and that the response patterns followed the explicit questionnaire routing (Filter Rules) the data did not always arrive in the expected format for edit and imputation. CANCEIS runs off a data dictionary where all variable ranges (categorical/numeric data) and patterns (alphanumeric data) are pre specified. If data did not arrive to the imputation process in the expected format then CANCEIS would fail, preventing any further progress in processing the data. Within DSP this was initially perceived as a bottleneck when actually the imputation process provided an invaluable service since without the CANCEIS data dictionary some of these errors may never have been detected. For example question marks were left in the postcode fields and there was a high level of inconsistency for questions that had tick-box and written responses such as the address fields.

15. The quality and assurance role the edit and imputation system played was a highly valuable process within the wider automated environment since errors detected after imputation would have been more difficult to correct due to the manual editing process. Errors of this kind, such as the inconsistencies in the address fields, caused observed data to be overwritten which reduced small populations, so the processes in DSP before edit and imputation had to be amended to correct the data. For example, the response for each address question was captured in several variables as depicted in Figure 3. There was a tick box response, a written in address, a postcode and a written in country each recorded in separate variables. The question contained implicit filters where an address or country was only required if the corresponding tick boxes were chosen. The valid combinations of values for the variables that were imputed are shown in Table 1.

Figure 3: Second Address Paper Collection
Source: 2011 England and Wales Questionnaire

Table 1: Table of valid response combinations

INDICATOR	POSTCODE	COUNTRY
1	NCR	NCR
2	Postcode	NCR
3	NCR	Country Code

NCR = No Code Required

16. Initially the Filter Rules for addresses did not edit the implicit filters and there was a high level of inconsistency between the tick box and written in fields. This affected the pass rates for the entire module, but also led to the address being overwritten during imputation since those with a second address formed a smaller population and there was a high chance of the donor not having one. The Filter Rules were amended so that records with an observed value in the postcode or country code fields would get the correct tick box response before imputation. This helped to maintain the smaller populations and also increased the donor pool for those populations.

B. Additional Editing

17. After amending the earlier processes, as expected, there were still complications with the imputation due to differences between the live data and the test data used during development. As a result the pass rates, especially for the demographics module, were lower than anticipated. A general strategy for minimal editing had been adopted during development in favour of allowing CANCEIS to edit and impute simultaneously. In principle it was thought to be better to edit in the imputation system where the underlying distributions could be taken into account rather than applying deterministic edits. In practice, due to erroneous response patterns, some editing had to be applied before CANCEIS and is described in the following sub-sections.

B.1. Relationship Editing

18. During the tuning process, with the early PUs there was a tendency for CANCEIS to fail imputation during the demographics module which contained the relationship matrix and all of the between person edit rules. The demographics imputation was particularly complex for households of sizes five and six where there was a higher number of relationship variables and a higher chance of an error being recorded. By improving the coherence of the relationship data, and thereby increasing the pass rate for the module, CANCEIS would have a higher chance of success. There was already an algorithm in place (RA1 in Figure 1) because the relationship matrix was known to be difficult to answer correctly. However, RA1 only corrected for the most common response errors where there was a strong indication of an error, for example where a parent or grandparent relationship had been recorded the wrong way round or where two records with the same parents were not recorded as siblings. To increase the demographic pass rates further, RA1 was adapted to apply deductive editing to the data using triangulation rules. For example, if person A was a parent of person B and person B was a parent of person C but the relationship between person A and C was missing then person A would be imputed as the grandparent of person C. While the triangulation rules did not correct observed inconsistencies, the values imputed were validated against the CANCEIS edit rules.

Table 2: Comparison of person 1 to 6 demographics pass rates before and after triangulation checks were added to RA1

	Before Triangulation Rules	After Triangulation Rules	Difference
Number of Households edited by RA1	4,000	10,701	6,701
%	1.78	4.77	2.99
Demographics Module Pass Rates (%)			
Household Size			
1	80.20	80.41	0.21
2	78.00	79.13	1.13
3	65.60	73.23	7.63
4	66.95	73.71	6.76
5	60.80	67.44	6.64
6	57.25	65.41	8.16
Overall Total	**74.43**	**77.27**	**2.84**

19. Using an example PU, Table 2 shows the number of households edited by RA1 before and after the triangulation rules were added and the subsequent demographics module pass rates for each household size. After RA1 was updated, the number of households edited by the algorithm increased by 6,700 records. This lead to an overall pass rate increase for the demographics module of 3% with the largest increase of 8% for six person households.

B.2. Deterministic Edits

20. A high level of non-response was also observed in the marital status question. There was a precedent for editing this variable as low response rates, particularly for children, had been observed in

the past and a deterministic edit was applied to edit marital status in the 2001 Census. The majority of records with a missing value for marital status were children under the age of 16, perhaps because some respondents felt the question was not applicable to children. The layout of the question in 2011, as shown in Figure 4, may also have contributed; with no category labelled 'single' and the appropriate 'never married...civil partnership' category sitting above the other columns of responses. The following edit rule was applied:

> If marital status is missing and the person is not a spouse or civil partner to any other household member then make their marital status 'never married'.

21. In the largest PU, this rule edited 64% of missing marital status values reducing the non-response rate from 6% down to only 2% for this variable.

Figure 4: Marital Status Paper Collection
Source: 2011 England and Wales Questionnaire


4 On 27 March 2011, what is your legal marital or same-sex civil partnership status?

- ☐ Never married and never registered a same-sex civil partnership
- ☐ Married
- ☐ Separated, but still legally married
- ☐ Divorced
- ☐ Widowed
- ☐ In a registered same-sex civil partnership
- ☐ Separated, but still legally in a same-sex civil partnership
- ☐ Formerly in a same-sex civil partnership which is now legally dissolved
- ☐ Surviving partner from a same-sex civil partnership

22. Despite anticipating the potential need for additional editing and running CANCEIS inside a SAS wrapper, due to the system being intergrated in DSP, changes to the SAS wrapper were difficult to implement in a short time frame. When designing an imputation system to be run inside an automated platform without accurate up-to-date test data, it is important to build a flexible editing structure into the system to allow for late adjustments before a large amount of data is processed through the system without the edits having been applied.

C. Tuning CANCEIS

23. Only once all the data issues had been resolved could the tuning process begin. Since CANCEIS is a statistical process dependent on the data, it was anticipated that when the live data was available, the edit and imputation system would need to be adjusted. Tuning CANCEIS involved two main parameter files; the distance models and weights for selecting the donors and the parameters governing the way the data was searched for donors. Due to CANCEIS being integrated into the DSP system, the aim was to find one set of parameters that could be used to impute every PU since any changes to the input files were difficult to implement.

24. CANCEIS splits its search for donors into stages and then builds a pool of potential donors from each stage containing the donors most similar to the record to be repaired. This similarity is based on the statistical distance between observed values in both the potential donor and the failed record. The donor pool is refined and reduced further by calculating the minimum number of changes needed to repair the failed record using the data from each donor in a way that satisfies the edit rules. Consequently, the final pool of potential donors contains records that are most similar to the record to be repaired and whose values, if used in the imputation, make the smallest amount of change to that record. The final donor is then randomly selected from that list. The user can define the number of records to consider in each stage of the search and the number of potential donors to retain in the first and refined donor pools. If only a small number of donors are retained in each stage, there is a possibility that none of the potential donors

has data that can be used to repair the failed record in a way that satisfies the edit rules. It can also be the case that far too many observed values in the failed record would need to be changed to meet those rules. In both situations the final potential donor pool would remain empty and the record would fail to impute. To reduce this risk, the donor search could be increased, retaining more potential donors and thus increasing the chance of finding an imputation action that satisfied all the edit rules with the minimum change required. However this increased the risk of a CANCEIS resource error where there was not enough memory to store all of the possible imputation actions [this resource issue is not present in later versions of CANCEIS]. This would cause the imputation to stop before completion and the entire PU would fail in DSP. Due to the way CANCEIS was integrated into DSP, all changes had to be kept to a minimum so a general set of parameters had to be found that were wide enough to reduce the number of unimputed records but narrow enough to reduce the risk of a resource error causing the whole PU to fail. After a few PUs had been processed through DSP, a general set of parameters could be found that balanced these two risks in most cases.

25. A similar process was undertaken where the imputation weights were tuned to find default settings. In CANCEIS each variable is given an imputation weight which defines how important that variable is in the imputation model. For example a variable with a lower imputation weight has a higher chance of being imputed over a variable with a high imputation weight. Tuning the weights was done by examining the quality indicators for the imputation, that is, the recovery of the underlying distributions in the data, the level of change to observed properties and the prevalence of soft edits. To a degree the tuning was also a balancing act between placing more weight on key variables and the need to maintain small populations within the data. Age was generally given the highest weight because it was considered to be more reliably captured as well as a good predictor for a lot of the other variables. However it was found that the weights given to variables that effected small populations had to be increased in order to prevent the system from changing observed characteristics for them. For example, the weight on the second address indicator had to be increased to help prevent a decrease in the proportion of persons with a second residence. This process took place outside of DSP and after each change the diagnostics were examined before applying the changes to the live system in DSP. This process took around seven working days and once the tuning was complete the same weights were used for all PUs.

IV Fallback Methods of Imputation

26. Despite tuning the system to the live data, not all PUs ran on a standard set of search parameters. If a PU failed due to a resource error, a unique set of parameters for that PU had to be found. In a few extreme cases, due to the complexity of the relationship matrix, the demographics module for the five and six person households had to be restricted to CANCEIS's narrowest settings, for example, out of every 2000 households, only one record with the smallest statistical distance from the failed record was retained in the refined donor pool. This reduced the stochastic element of CANCEIS turning it into a deterministic process. However by comparing the diagnostics from these restricted settings with those from the default settings, it was found that the impact on the quality indicators of the imputation was minimal. In rare cases, even limiting the settings in this way was not enough to prevent a resource error from occurring. For these PUs, the record causing the error was isolated and imputed manually for the demographics module outside of DSP, sometimes changing as little as one or two relationship variables. The partially imputed record would then be updated on the database before imputation, allowing the rest of the record to be imputed in DSP. Around 20% of PUs required a unique set of search parameters and only four PUs contained a record that required manual imputation to be applied before processing in CANCEIS.

27. Even after the tuning process, and the potential for unique parameter files, not all records imputed in DSP leaving on average 157 people per PU unimputed. These failures were spread fairly evenly across all of the PUs with the majority of people being from six person households or larger. Since the person 7 plus imputation was dependent on the person 1 to 6 instance being successful, if the first six people failed imputation, the continuation people for that household were not processed through DSP. These records, along with any unimputed records, had to be imputed outside of DSP in an Ad Hoc environment; a secure computer network developed to safely interact with the core database. The majority of these records could be imputed using the same methodology as applied in DSP by making slight adjustments to the parameter files which altered the way CANCEIS searched for donors. However

in a minority of cases, even considering every other record in the file as a donor, some records could not be entirely be imputed jointly. Two methods of fallback imputation were used for these records; individual imputation and manual imputation. For individual imputation, some or all of the people in the household would be imputed as a single person household by CANCEIS for one module, usually the labour market module where there were no between person consistency rules. However in many cases one or two variables would be imputed manually before imputing the remaining variables jointly using CANCEIS in the usual way as shown in Figure 1. Fallback imputation using these methods was very infrequent and was only required for around 200 person records from the entire population (0.001%). Whether a PU required a unique set of parameters or a record was imputed in DSP or the Ad Hoc, for the vast majority of records the methodology was the same.

V Evaluation

28. At the time of writing, census processing is ongoing in preparation for formal release and therefore it is too early to provide any detailed quantitative analysis of outputs from the imputation system. However, preliminary analysis indicates that in general the system worked well. The extensive testing carried out at the beginning of the census project provided assurance that the methodological principles implemented in CANCEIS would serve as an appropriate foundation for the 2011 edit and imputation system. This was confirmed by the diagnostic tools built into the system which were designed to compare the distributional properties of the data pre and post-imputation. (A more detailed overview of these diagnostic tools is presented in Wardman, Aldrich & Rogers, 2012).

29. Table 3 shows several comparisons from the 2001 Census EDIS system and the 2011 edit and imputation system. There were around 4.1 million additional person records processed in 2011 and the proportion of people requiring imputation was 7% higher than in 2001. However this is partly due to the strategy of editing and imputing simultaneously compared with 2001 where the data was edited prior to imputation. Despite this the average processing time for a PU in 2011 was up to four times faster than in 2001 indicating an overall gain in processing efficiency.

30. Methodologically, there were 82% of people jointly imputed with their household in 2011 compared to only 34% in 2001. The remaining 18% of people that were imputed individually in 2011 were mainly single person households, continuation people and communal people for whom an individual imputation was planned, with only 16 people being imputed as an individual when a joint imputation was attempted. Despite an increase in the number of people requiring imputation, the fallback imputation methods in 2011 were used for almost thirty times fewer records than in 2001. Since the majority of records that were imputed using the fallback methods in 2011 were imputed using CANCEIS in the Ad Hoc environment, only around 200 people were imputed using more than one method compared to over one million people in 2001, thus maintaining consistency in the methodology.

Table 3: General comparisons between the 2001 and 2011 Census imputation systems

	EDIS: 2001[a]	SAS/CANCEIS:2011[b]
Person records passed through imputation system	49.4 million	53.5 million
Persons needing at least one variable imputed	13.8 million (28%)	18.6 million (35%)
Average number of records in a Delivery Group	500,000	530,000
Average time to impute a Delivery Group	48 hours	12 hours
Persons imputed taking into account multivariate joint distributions between variables (% of those imputed)	34%	82%
Persons imputed as individuals	72%	18%
Persons imputed using alternative methods to that implemented in primary imputation system	3%	0.10%
Persons imputed by more than once method	Over 1 million	200 approx.

a. Census 2001 Review and Evaluation (ONS, 2003)

b. Data derived through the Census 2011 SAS/CANCEIS system diagnostics

31. Overall the planned methodology worked well in achieving a complete and consistent database with only a few minor changes required in response to the observed data once processing had begun. Using CANCEIS as the foundation for the imputation system increased processing efficiency and allowed flexibility when handling both the PUs and the records that failed to impute in DSP enabling consistency throughout the imputation process. The inclusion of the relationship algorithms were vital in stabilising the data and increasing the donor pool, while the tuning process was crucial in successfully processing the majority of PUs using a general set of input parameters and reducing the reliance on the fallback imputation methods. There were some limitations to the edit and imputation strategy however these mostly came from external factors, for example, integration in DSP which reduced interaction with the system and the lack of suitable test data leading to last minute changes to the process.

32. Leading on from these initial results, as part of the evaluation process, there is an extensive list of post-processing research planned to analyse the finer details of the imputed data and the imputation system overall. Quantification of discrete edit failures, imputation variance, and modal effects based on internet/paper capture are just three high-priority examples from that list.

VI. References

Wardman, L., Aldrich, S., Rogers, S. (2012): Item imputation of Census data in an automated production environment; advantages, disadvantages and diagnostics. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Oslo 24-26 September 2012.

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

ONS (2003): Census 2001 Review and Evaluation, Edit and Imputation: Evaluation Report.

Wagstaff, H.F. (2011): 2011 UK Census: An Overview of the Edit and Imputation Process. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Ljubljana, 9-11 May 2011.

CANCEIS Development Team (2009): CANCEIS Version 4.5 User Guide. Social Survey Methods Division, Statistics Canada, February, 2009.