WP3.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

2011 UK CENSUS: AN OVERVIEW OF THE EDIT AND IMPUTATION PROCESS

Invited Paper

Prepared by Heather Wagstaff, Office for National Statistics, United Kingdom[1]

I. Introduction

1. In 2001, the UK Census Offices processed about 27 million returns which contained responses for almost 60 million people. Overall, some 28 per cent of the responses contained one or more erroneous items. Despite significant improvements in questionnaire design, including the introduction of an internet questionnaire as an additional mode of collection, the 2009 Census Rehearsal has indicated that similar levels of erroneous responses may be present in the 2011 Census.

2. This paper outlines the development of the 2011 UK Census editing process. At the time of writing, the Census fieldwork is nearing completion and data capture and coding has commenced. Section II of the paper provides background information on the 2011 UK Census Editing Strategy, an overview of CANCEIS, and briefly describes the 2011 UK Census processing operation. Section III describes implementation of the editing process including methods to correct erroneous responses to the relationship matrix. Section III also describes the treatment of statistical edit rules and the process of imputing skeletal records as part of the coverage adjustment process. Finally, Section IV closes with some concluding remarks.

II. Background

A. UK Editing Strategy

3. The 2011 UK Editing Strategy was developed with the primary aim of imputing for all item level missingness and resolving inconsistencies in the responses for the households and persons affected. The Strategy is driven by three key principles:

1. all changes 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 in order to provide a complete and consistent database.

4. In addition, a number of broad principles were followed when developing the internet questionnaire in order to ensure a high quality design:

1. no unnecessary changes were made from the paper questionnaire;
2. respondent burden was minimised by making the user experience for internet data collection (IDC) as simple and easy as possible;
3. adherence to web design best practice where possible;

[1] Heather.Wagstaff@ons.gov.uk

4. ensuring accessibility for everyone including those who require assistive technologies when using the Internet.

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

B. Overview of CANCEIS

6. For 2011, CANCEIS forms the cornerstone of the UK Editing Strategy. 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, and then the imputation is performed by searching for donors. Changing the order of the operations in NIM allows CANCEIS to solve larger and more complex edit and imputation problems (Felligi-Holt, 1976; Bankier, 2000).

C. Overview of the 2011 Census Process

7. For the first time in the UK, the 2011 Census has employed a mail out/back enumeration method which was made possible by the development of a central Address Register. About 95% of the questionnaires were mailed out to residential addresses. The remainder were hand delivered in areas known to be hard to enumerate and also to Communal Establishments (Collectives). In addition, all households were offered the option to complete the questionnaire on-line. The Address Register has been central to identifying non-responders and to targeting the follow-up field force efficiently and effectively.

8. The 2001 Census estimated that about 99.5% of the household population lived in households containing 1 to 6 people (ONS, 2003). This group of respondents are known as the 'main household population', with households containing seven or more people known as 'large households'. The 2011 paper questionnaires were designed to align with these population groups, i.e. the main household form contains responses for 1 to 6 people, whilst large households are required to complete an additional one or more continuation forms. Whilst all residential address were mailed a single household form, those containing seven or more persons were asked to contact the Census Helpline to request one or more continuation forms. However, the Internet questionnaire collects information for up to 30 people which is the largest household size for which tabular outputs are produced. There is concern that larger households completing paper questionnaires may be under represented, hence leading to a clear modal effect between collection methods (Wagstaff and Wallis, 2008, Wagstaff and Dalton, 2009).

9. Similarly to 2001, the 2011 data capture and coding operation has been outsourced to Lockheed Martin. The forms are scanned and captured using a combination of OMR and OCR. Characters recognised with a low degree of confidence are sent to manual key repair and a set of preliminary edits are applied to ensure the validity of the output data. The Statistics Canada coding tool ACTR (Automated Coding using Text Recognition), is applied for complex coding to convert textual responses to numerical format. Responses which could not be coded automatically, or by computer assistance, are sent to expert coders to assign the appropriate code. An automatic quality assurance system is integrated within the capture and coding sub-systems to assess whether the coding is consistent and meets with pre-specified quality standards. The data will be transmitted to ONS electronically before loading to an Oracle database. Once loaded, the data will pass through a series of essential validation steps which facilitate the operation of the key statistical processes (Edit and Imputation, Coverage Assessment and Adjustment and Statistical Disclosure Control).

However, two of the validation processes have a significant impact on the structure of the entities (Households and Communal Establishments):

1. Remove False Persons: which identifies and removes person records where there is insufficient evidence to suggest that a person exists; the majority of false persons are generated by recognition error at capture;
2. Reconcile Multiples: which links all valid responses for a single entity and removes duplicate responses.

10. Both processes apply complex algorithms and generate the need for further validation to ensure the integrity of the final entities. The procedure, known as Post Remove Multiples (RM) Validation includes: truncating household size to 30 persons; counting the number of valid persons within each entity; renumbering person records from 1 to n (where n = person count); renumbering relationship responses to realign with the revised person numbering; resolving inconsistencies generated by 1 and 2 above (where a person's response refers to person 1 but the original person 1 has been removed); setting indicator flags to identify which persons have been amended; and running diagnostic reports. It is important to note that no households or persons are deleted from the Census database. Duplicate records, spurious persons and persons numbered 31 or above are simply disregarded from the logical data model.

III. 2011 UK Census Editing Methodology

11. The editing methodology has been developed over a number of years to ensure that it is both robust and efficient. It was developed in consultation with key stakeholders and subject matter experts including representatives from each of the UK countries, as well as demographers. In addition, the general approach to all methodology has been quality assured by academics from Southampton University and formally endorsed by the UK Census Methodology Advisory Committee.

12. The 2011 data will be processed as a series of 104 processing blocks, formed from clusters of contiguous Administrative Areas, known as Delivery Groups (DG's), each containing about 500K population. Editing the Census data is necessarily complex due to the multiplicity of household living arrangements, the complexity of the question set and the necessity to complete the process within very tight timescales. Early research indicated that the optimal approach to editing was to partition the data within each DG into homogeneous population subgroups, or editing modules, which align with the census questionnaires (main household population, large households and collectives).

13. The 2011 question set contains a number of new and revised questions. There are 8 household questions, an inter-person relationship matrix and 43 person questions (5 of the latter relate solely to routing). The complex question set has a significant impact on editing, for example, the 43 person questions generate 148 different response patterns.. Hence, it is clear that a fully joint imputation approach is infeasible for a question set with this level of complexity. Therefore, it is statistically efficient to partition the person data, within the editing modules, and apply joint imputation to the partitions. This approach is also operationally efficient: where responses to routing questions are missing it is not possible to determine whether responses to subsequent questions should be present or not. For example, age must be present to indicate whether there should be responses to the Labour Market questions. The partitioning approach also serves to maximise the available donor pool.

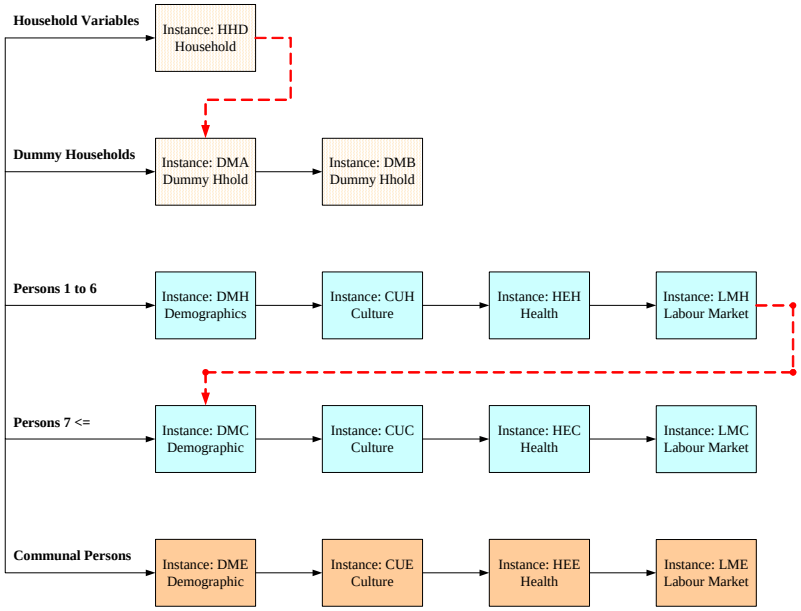
14. The next issue was how best to form the partitions. The highest priority variables for the Census Outputs are those that support the population count and formed the basis of the primary partition. So, using a sample of 2001 data, standard logistic regression techniques were applied to the variables in order to identify their strongest predictors. Findings indicated that the variables relating to *age, sex, marital status, activity last week* and *relationship to household reference person* (HRP) formed a self predicting set (i.e. were strong predictors for one another) and formed the basis of the primary partition ('Demographics'). Including *activity last we*ek, from Labour Market, in the Demographics module strengthened the partitioning approach because it negates the likelihood of imputing *age* as less than 16 years where the respondent reported an '*activity*' which indicates they are not a child. The remaining partitions were distinguished by the position of two main routing questions relating to: (1) students living away from home on Census night; and (2) Labour Market

questions. Questions which immediately follow the student filter cover two main topics relating to Culture and Health. Therefore, each person level module was divided into four partitions broadly relating to: Demographics, Culture, Health and Labour Market.

15. A high level diagram of the prototype editing process is shown at Figure 1. In this configuration, the data are fully edited by CANCEIS which is contained within a Java wrapper to allow full integration with the Census Downstream Processing (DSP) operation. Figure 1 shows that there are five modules: Household Variables, Dummy Forms, Household Persons 1to6, Household Persons 7 and over; and Communal Persons (those living in Collectives).

16. Dummy Forms relate to unoccupied dwellings which act as non-response indicators, in the Census Database and the Address Register, to confirm that the dwellings exist. Thus, the indicators contribute to the overall dwelling count. Dummy Forms contain a subset of the household question set: those which the enumerator completes by observation; together with the reason the dwelling is unoccupied. There is a requirement to impute the Dummy Forms as their geographical locations are assigned to synthetic households which are estimated to have been missed by the Census. The Dummy Households are imputed in two stages: (1) using clean household records as donors for the subset of household questions; and (2) using other Dummy Households as donors where the reason for the dwelling being unoccupied is missing.

Figure 1: Prototype 1 - 2011 Census Editing and Imputation Process



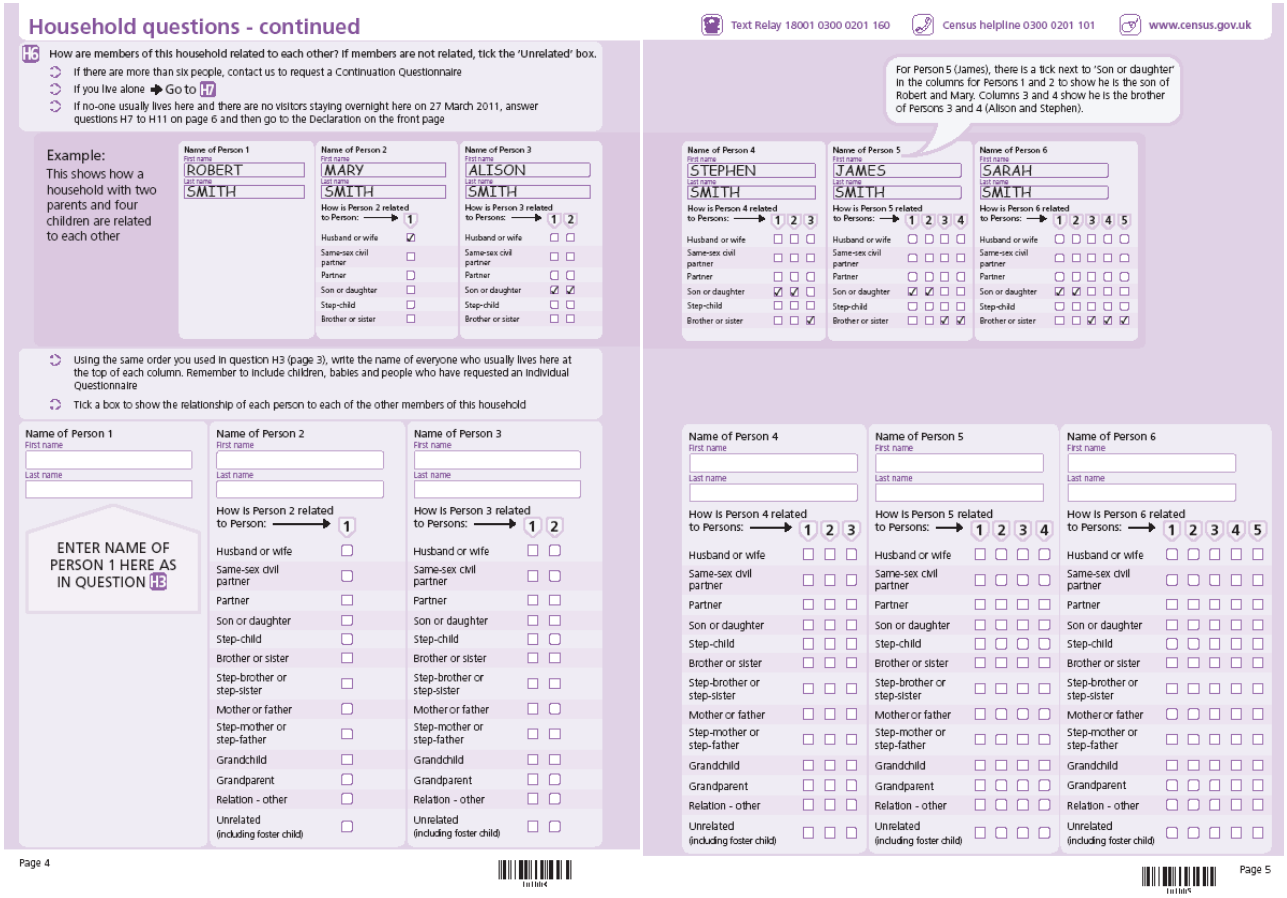
17. Household persons are processed by separate strata for households sizes 1 to 5 together with a '6+' stratum which contains households of size 6, together with the first 6 people from larger households. The upper limit of 6 persons is applied to align with the main household population but also because the failure rate for larger households would be too high. Hence, persons 7 and over are processed in an overflow stratum where only the within person edits are applied. The imputed relationship variables for persons 1to6, together with the key demographic variables of the HRP, are carried forward to the Persons 7+ module but are not imputable. There are insufficient large households to assign to individual strata. Such an approach would, at best, result in an unacceptable level of donor reuse causing disturbance to distributional accuracy and, at worst, result in an unmanageable failure rate. Further, in the reality of live running, the 'nearest neighbour' donor households are unlikely to be statistically 'near' to the recipient.

18. Development of the editing process went well whilst the only relationship included in the Demographics module was to the HRP. As other relationships were introduced, a negative impact on the overall quality of the imputed data was observed, especially amongst the key demographics. To describe the complexity of the issue: Section A provides background information about the relationship matrix; its role in deriving household composition; and the main findings from researching the problems. Section B describes the general approach to resolving the issues.

A. Overview of 2011 Census Relationship Matrix

19. The 2011 Census relationship matrix was specifically designed to meet United Nations and Eurostat recommendations to identify and classify family units within households. It was also designed to meet user needs to identify 'hidden' and 'reconstituted' families. The Household Composition Algorithm (HCA) is a complex derived variable, based on the relationship matrix, which identifies and classifies family units within households. Hence, the HCA must account for the multiplicity of modern day living arrangements and its accuracy is dependent on the accuracy and consistency of the relationship matrix. The matrix is the most complex question to complete and as household size increases so the number of relationship responses increase (Wagstaff and Wardman, 2009; Wagstaff and Dalton, 2009).

Figure 2: Relationship Matrix – Paper Collection
Source: 2011 England and Wales



20. When completing the paper questionnaire, the form filler must record the names of the household members three times in: (1) the listing grid; (2) the relationship matrix; and (3) the person questions. The relationship matrix occupies a double-page spread which is constructed to collect the relationship of every household member to the household reference person and to the persons ordered before the respondent in the matrix. See Figure 2 above. The instructions are lengthy and some respondents find the layout confusing. It has not been possible to analyse 2009 Rehearsal data but previous analysis of a sample of 2001 records found inconsistencies amongst person ordering. Specifically, cases existed where the ordering of persons in the relationship matrix differed to the person questions. This was especially marked amongst the larger households. Also, a number of relationships were reported the wrong way round, for example selecting ‘parent’ instead of ‘child’, or ‘grand-parent’ instead of ‘grand-child’. Some 6.7% of households returned at least one multi-ticked response to the relationship matrix of which 86.5% of the multi-ticks were irresolvable and passed to imputation for resolution. The research indicated that the propensity for multi-ticked responses increased with household size and ethnicity (Fearn, Rogers and Wagstaff, 2007). Subsequent research also identified a clear break in data quality at the Household/ Continuation form threshold which was especially marked in the quality of relationship information.

21. The Internet questionnaire makes use of personalisation techniques which serve to mitigate respondent error. The form filler is asked to record the number of usual residents and write their names in the listing grid. The electronic interface then creates the correct number of sets of individual questions and populates the names in the relationship matrix and individual questions. The relationship question is then formulated, "How is John Smith related to Susan Smith" and will hopefully negate the majority of response errors. However, if large households are related to poor literacy, for example, then it may be that respondents become confused and frustrated by the volume and complexity of the on-line relationship questions, break-off early in the process and not realise the requirement to complete the questionnaire. However, overall, we expect to observe an element of modal variation between the collection instruments and especially amongst relationships.

B. Resolving issues with relationship matrix

22. Where erroneous relationships existed, they had unexpected influence on other directly observed responses. For example, the prototype would regularly change a directly observed age, in preference to relationship, even though age was assigned the highest weight. Therefore, it was crucial to amend erroneous relationship responses before the data passed to CANCEIS.

23. The edit rules for the relationship matrix are termed 'conditions'. Overall, 8 conditions were identified where, when taken in combination with other variables, it was clear that the form-filler had misunderstood the question. Examples include: parents and children reported the wrong way round; similarly for grand-parents and grand-children; and children with two parents in common who were not reported as siblings. These 8 conditions formed the basis of an algorithm which was developed to be applied to all persons in the household. Thus, the purpose of the relationship algorithm (RA) is to identify and 'correct' inconsistencies within the data. It is run prior to imputation, thus maximising the number of potential donors with consistent data. The algorithm only switches relationships in cases of certainty and does not impute for missing or other erroneous values.

24. A second algorithm was developed for large households which was necessary because the Persons 1to6 module includes the first 6 people from large households. This algorithm ensures that the higher order people do not have an erroneous relationship to the HRP. The data is then passed to CANCEIS for the Persons 7+ module before a final algorithm is run to ensure that any remaining inconsistencies or missing values are resolved between the relationships of persons 7 and over. A set of indicator flags are set to identify which conditions have been invoked for each respondent and by which algorithm. The method has been thoroughly evaluated to ensure that it is robust and maintains distributional accuracy. The method has been quality assured by leading academics from Southampton University.

25. The approach was applied to a set of test data from a large English inner city conurbation which was selected for its high variability in terms of both ethnicity and household structure. The dataset contained 13,877 households, and 94,432 persons, in households of sizes 7 to 21. The data were then processed through the second prototype imputation system in 5 stages as follows:

Stage 1 Apply Relationship Algorithm (RA) 1 to ALL household persons to correct for mis-reported relationships.

Stage 2 Impute the Persons 1to6 from applying the complete set of edit rules.

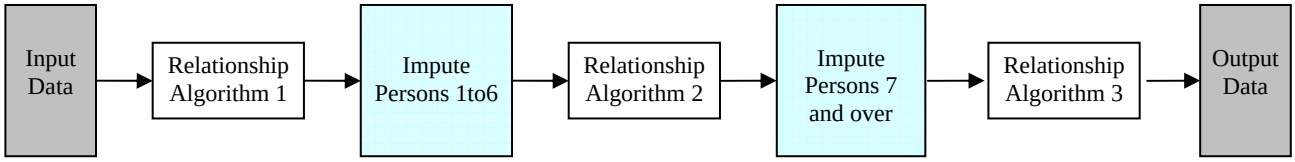
Stage 3 Apply RA2 to persons 7 and over to ensure that the imputation has not introduced erroneous relationships to person 1;

Stage 4 Impute Persons 7 and over applying within person edit rules only;

Stage 5 Apply RA3 to impute missing values in relationships for persons 8 and over (person 7 only completes Relationship to HRP).

26. The process is shown diagrammatically in Figure 3.

Figure 3: Prototype II: ensuring consistency of the Relationship Matrix



27. The effectiveness of the approach is demonstrated by the evaluation of the proportion of failures for 6 of the relationship conditions applied by RA1 and RA2:

Age Parent:	A person aged less than 12 cannot be a parent.
Parent Age:	A parent cannot be less than 12 years older than their child.
Age S. Parent:	A person aged less than 12 cannot be a partner or step-parent unless their country of birth is outside of the UK.
One Spouse:	A person cannot have more than one spouse or civil partner.
Spouse Partner:	A person with marital status of single must not have a spouse or civil partner.
Triangulation:	Two people with at least one parent in common cannot be married or civil partners or partners with each other

28. The data were analysed by scrutinising the outcome for Persons 1to6 and Persons 7+ separately. The household level failure rates are shown in Table 1. For example, amongst Persons 1to6, some 22.28% (3092/13,877) of households failed one or more of the six conditions, similarly 16.84% (2337/13,877) for Persons 7+. As expected, Parent Age was the highest failing condition in this group accounting for just over 78% (2,419/3092) of the failing households for Persons 1to6 but only 20% (476/2,337) for Persons 7+. RA1 cleared almost 30% (906/3092) of the erroneous households for Persons 1to6 and almost 87% (2029/2337) for Persons 7+. As expected, CANCEIS output fully complete and consistent data for Persons 1to6.

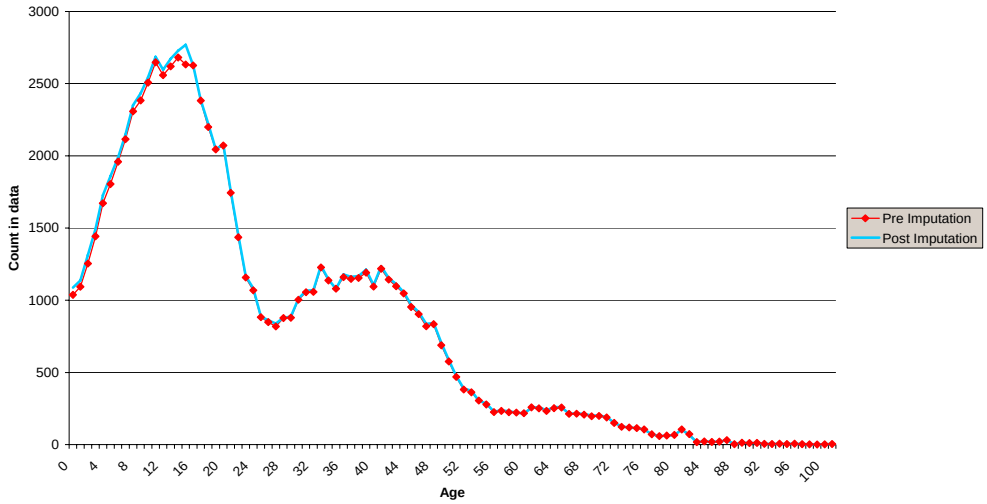
Table 1: Household level failure rates

Edit Rule	Persons 1 to 6				Persons 7 to 21							
	Raw Data		Post P1to30 Algorithm		Raw Data		Post P1to30 Algorithm		Post P1to6 Imputation		Post P7+ Imputation	
	n	%	n	%	n	%	n	%	n	%	n	%
Age Parent	1,210	8.72	742	5.35	345	2.49	188	1.35	127	0.92	85	0.61
Parent Age	2,419	17.43	1,644	11.85	476	3.43	275	1.98	221	1.59	110	0.79
Age Partner Stepparent	179	1.29	144	1.04	21	0.15	19	0.14	19	0.14	20	0.14
One Spouse	233	1.68	233	1.68	0	0.00	0	0.00	46	0.33	1	0.01
Spouse Partner	65	0.47	65	0.47	8	0.06	0	0.00	4	0.03	0	0.00
Triangualtion	708	5.10	506	3.65	35	0.25	24	0.17	23	0.17	0	0.00
Total Failing H'holds	**3,092**	**22.28**	**2,186**	**15.75**	**2,337**	**16.84**	**308**	**2.22**	**295**	**2.13**	**126**	**0.91**
Total Passed H'holds	**10,785**	**77.72**	**11,691**	**84.25**	**11,540**	**83.16**	**13569**	**97.78**	**13,582**	**97.87**	**13751**	**99.09**

29. Prior to applying RA2, Persons 7to21 contained 2.13% (295/13,877) of households with erroneous relationships, of which 57.2% (169/295) were amended by the algorithm. The Persons 7+ imputation module does not correct for relationships other than to the HRP but any outstanding erroneous values are imputed by RA3. At the end of RA2, the process had amended 95.92% (2,966/3,092) of the households which contained erroneous relationships. Conversely, just 4.08% (126/3,092) remained for correction by RA3.

30. Finally, the distributional accuracy of the key variables *age*, *sex*, *marital status*, *activity last week*, *relationship to HRP*, was evaluated pre to post imputation. The evidence suggested that the imputation process maintained univariate and bivariate distributions for each five of the variables. As an example, Figure 4 shows the *age* distribution before and after imputation.

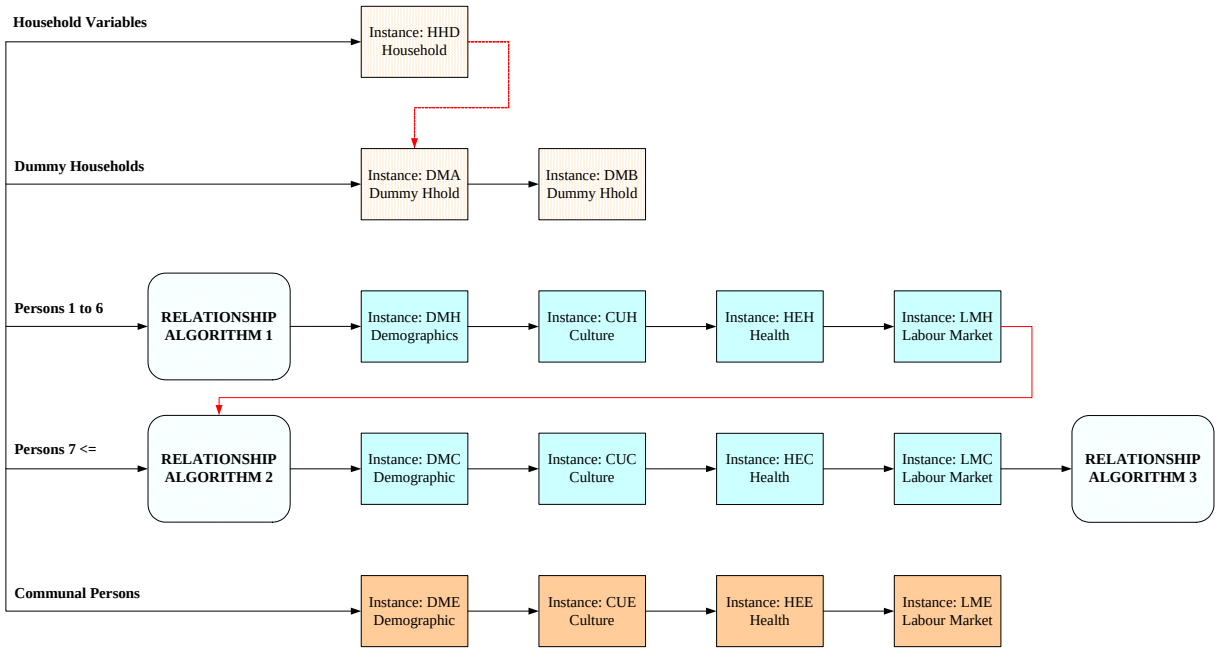
Figure 4: Distribution of AGE – pre vs. post imputation



C. Final 2011 Census editing process

30. Following the development of the relationship algorithms the final editing process was reconfigured to embed the three algorithms as shown in Figure 5. The process will be implemented as a series of five modules with dependencies between the (1) Household module and the Dummy Household modules; and (2) Persons 1to6 and Persons 7+. There are no dependencies for Communal Persons. Therefore, it is possible to run the process as three parallel streams. RA1 will be applied prior to Persons 1to6 module. Once the Persons 1to6 module is complete, RA2 is applied to persons 7 and over to ensure that the imputation has not introduced erroneous relationships to the HRP. Finally, once the Persons 7+ module is complete, RA3 is applied to correct any remaining erroneous relationship values amongst persons 7 to 30.

Figure 5: 2011 UK Editing Process



D. Post Coverage Item Imputation

31. The aim of the 2001 Census Coverage Adjustment Strategy was to create a census database that was fully adjusted for the under enumeration at both the household and person level for all census outputs. 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 calibrated to the estimates at the Administrative Area level. The weights were then applied for the selection of donor households and people estimated to have been missed from counted households. Each imputed household was placed into either a Dummy Household, an empty household, or into an Output Area within the Administrative Area; individuals who were missed from responding households were imputed into an existing household. A final stage of the process adjusted the post-imputation Census database to ensure that the targets for household size and age-sex estimates were met exactly within each Area. (Steele et al, 2002).

32. The 2001 coverage imputation process copied records that had been previously treated by the item imputation system and hence, were known to satisfy the edits. However, creating exact copies of existing records onto the database did not necessarily create a database that truly reflected the heterogeneity in the population. In addition, the imputation models only controlled a subset set of characteristics and the system relied on the inter-relationships between variables within records to control the other characteristics. Therefore, there were concerns that there was a lack of heterogeneity in the uncontrolled variables, for example, occupation. Additionally, the donor search was constrained geographically to try to avoid creating implausible outcomes, e.g. travel to work patterns, which then restricted the records available for imputation. (Sexton and Brown, 2010)

33. In 2011, the Coverage Adjustment process will follow a similar, but refined, process to that for 2001. The process will estimate the number and geographical location of households and people missed by the Census and place a set of 'skeletal' records into Dummy and empty dwellings on the Census database. The skeletal records will contain the key demographic variables which were controlled by the adjustment process. CANCEIS will then be applied to impute for the missing items whilst holding the controlled variables fixed. At the high level, the editing process is similar to that shown in Figure 4 but without the imputation of Dummy Forms and RA1 (since all relationships will be consistent or missing). Early results are encouraging but the quality of the imputation process will be dependent on the proportion of missing households.

IV. Concluding Remarks

34. ONS will start to receive the live 2011 Census data from Lockheed Martin in June. Whilst the editing process has been fully specified the Methodologists are currently engaged in a significant amount of customer testing of the integrated process. However, the main focus over the coming months will be on tuning CANCEIS and quality assuring the full set of DLTs and relationship algorithms.

35. Applying CANCEIS as the cornerstone of the 2011 UK Census Editing Strategy has lead to significant cost savings and efficiency gains. The flexibility of the parameter driven system has negated the difficulties associated with the variance between the question sets of the four UK Countries. The CANCEIS data dictionary has offered a great deal of transparency to the process and facilitated late changes to complex coding frames and filter rules. However, there is still a large amount of work to do to ensure readiness of the editing process for the 2011 Census.

V. References

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

Fearn, V., Rogers, R., Wagstaff, H.F., (2007) 'Towards the 2011 UK Census Edit Strategy: Resolving responses to the relationship question". Paper presented at GSS Methodology Conference, London, June 2007. Available at www.statistics.gov.uk/events/gss2006/downloads/C3 Fearn.doc

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.

Steele, F., Brown, J. and Chambers, R. (2002) "A controlled donor imputation system for a one-number census," Journal of the Royal Statistical Society A, 165. 495-522.

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

Wagstaff, H.F. and Wallis, R. (2008) “First thoughts on editing in mixed modes in the 2011 England and Wales Census’. Working Paper 4. UN/ECE Work Session on Statistical Data Editing, Vienna.

Wagstaff, H.F. and Dalton, S. (2009) “Editing in mixed modes in the 2011 England and Wales Census”. Working Paper 14. UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Wagstaff, H.F. and Wardman, L. (2009), “Edit and Imputation of the 2011 UK Census”, Working Paper 17, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.