WP.37
ENGLISH ONLY

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) Censuses

CANADIAN CENSUS E&I – LESSONS LEARNED FROM 2006 WITH PLANS FOR 2011

Invited Paper

Prepared by Michael Bankier, Statistics Canada

I. INTRODUCTION

1. Edit and Imputation (E&I) of the 2006 Canadian Census was successfully completed in the fall of 2007. The Canadian Census Edit and Imputation System (CANCEIS) was used to perform all deterministic and donor imputation. Section II lists the major changes that were made to how collection and processing was done for the 2006 Census. Section III outlines the impact of the coverage edits which cause persons to be added or subtracted from questionnaires. Section IV describes how the occupancy status of dwellings was verified and then how imputation of total non-response households was carried out. Section V shows the sequence in which variables had E&I applied and why it was not possible to impute them all simultaneously. Section VI goes over the process used to develop the E&I software in a timely and effective manner. Section VII describes the imputation of the demographic variables with a focus on age. Section VIII suggests some possible enhancements being considered for the 2011 Census. Finally, Section IX provides some concluding remarks.

II. A SUMMARY OF MAJOR CHANGES MADE TO THE 2006 CENSUS

2. Major changes were made to how field operations and data processing operations were carried out in the 2006 Census compared to the 2001 Census. These are summarized below.

3. Previous censuses had enumerators list dwellings and then leave a questionnaire which was completed and mailed back by the respondent. For 2006, an existing address register was updated by field work for the more urbanized areas prior to the census and then questionnaires were mailed out to nearly 70% of the households in Canada from the resulting master list of dwellings.

4. Eighty percent of the households received a short questionnaire while twenty percent received a long questionnaire (these will also be called short/long forms). Households in more remote areas (approximately 3% of the total number of households) were interviewed in person using a long form. On the long questionnaire, respondents had the option of giving permission to link to their tax form rather than answer questions on income. Approximately 85% of the respondents gave permission to do so.

5. For the first time in 2006, respondents could respond via the Internet and 18% of the households did so. To use the Internet option, the respondent had to enter an Internet access code provided on their paper questionnaire. The level of encryption used with the Internet application was higher than that typically used with on-line banking transactions. The amount of follow-up required because of incomplete responses was much less for Internet questionnaires than for paper questionnaires.

6. In 2001, the questionnaires were keyed while in 2006 they were scanned and then data captured using automated intelligent character recognition with some manual keying.

7. In 2001, local enumerators followed up for total non-response households while in 2006 this was done by specially trained staff operating out of 36 Local Census Offices. This process was called non-response follow-up (NRFU).

8. In 2001, local enumerators followed up for partial non-response while for 2006 this was done using computer-assisted telephone follow-up from three call centres. Some 10.3% of households that returned a questionnaire required this failed edit follow-up (FEFU).

9. These new approaches allowed the number of field staff required to be reduced by 46% for 2006. Because of widespread labour shortages in some parts of the country, however, the collection period had to be extended from mid-July to the end of August. Follow-up had to be focused on problem areas and some staff from head office had to be deployed to the regions to help with NRFU. In the end, a 2.8% non-response rate was achieved at the Canada level in 2006 compared to a 1.6% non-response rate in 2001.

III. IMPACT OF COVERAGE EDITS[1]

10. In the 2006 Census, the coverage edits process was designed to deal with inconsistencies related to the number of usual residents in both private and collective households. After initial scanning at the Data Processing Centre (DPC), coverage-related edits were applied to each household by an automated system. Questionnaires failing these edits were sent to the interactive coverage edit resolution system. There, operators corrected these problems, referring difficult cases to subject matter experts. Operators could take several different actions to resolve edit failures, notably deleting or adding persons, merging one person's data into another, removing persons from a household but retaining them for possible linkage to another household or deleting forms. Several fields on the forms could also be modified, particularly for collectives.

11. A total of 645,216 people (2.04% in terms of the published population count of 31,612,897) were deleted from private households, while 43,485 persons (0.14% in terms of the published population count) were added. People were deleted for many reasons, the most significant being that many people who did not exist were captured during the scanning process (because of errors such as stray marks being captured as responses to person-level questions, or respondents entering data into a blank person column by mistake) and a large number of duplicate responses were received that had to be deleted. In 85% of these cases, one of the duplicate responses received was as a result of NRFU.

12. Most people that were added to private households were missing because, while the names of up to ten persons could be listed, there was only space to provide data for six persons on the short form and five persons on the long form. The respondent was supposed to request an additional census form if required but often they did not. This problem was exacerbated, compared to 2001, by the reduction in the size of the long form from six to five people in 2006.

13. In addition, more than 18,000 persons were deleted from collective households, and another 22,042 were added. Most people deleted were temporary residents, duplicates or false persons due to capture error. Added persons were mostly those for which the enumerator was unable to obtain a response for some reason, many in total non-response dwellings.

14. The results of quality assurance monitoring performed during production showed that operators generally resolved the edit failures correctly, with estimated accuracy rates of 97% for private households and 85% for collective households. The errors that were made did not have a significant impact on over

[1] The information in this section is taken from the executive summary of the report "2006 Census RIVT Coverage Edits Evaluation" dated May 24, 2007 by Steve Rathwell and Glenn Hui of Statistics Canada.

or under-coverage. The number of persons deleted, added, not deleted, or not added incorrectly was quite small; each of these four types of error accounted for less than 0.02% of all persons enumerated.

15. Other processes that have a direct impact on coverage are FEFU and the multi-link edits. In FEFU, 41,295 persons were added and 22,285 persons were deleted. Persons were added mainly to households that had incorrectly indicated that all residents were temporary or foreign, or had indicated that they were unsure whether to include one or more residents. Many of the persons deleted were also in these types of households. About 36% of the persons deleted in FEFU had been added in resolution of coverage edits (about 19% of all persons added there).

16. Multi-link edits are designed to ensure that all forms for the same household are processed together, and to identify blank or redundant responses where possible. In multi-link edits, forms were deleted if they were received after responses for the household had already been written to the census data base, or after the household had been contacted in FEFU, or after a Census Help Line (CHL) response had already been received for the same household. About 126,000 questionnaires (in 120,000 households) were deleted for this reason. This was done in an automated fashion, the assumption being that these forms were duplicates of the responses received earlier. Analysis done to date indicates that approximately one third of the forms deleted contained people that were not included on the forms that were processed for those households. About 60% of the households with a form deleted by the multi-link edits had at least one NRFU response (as either the response received initially or the response deleted).

IV. IMPUTATION OF TOTAL NON-RESPONSE HOUSEHOLDS[2]

17. The Dwelling Classification Survey (DCS) was used to estimate the error rates in classifying dwellings in the self-enumerated collection areas as occupied or unoccupied in the field. Based on this information, adjustments were made to the census data base. The DCS selected a random sample of 1,405 self-enumerated Collection Units (CUs - there are 50,782 CUs in Canada) which were revisited in July and August 2006 to reassess the occupancy status as of census day for each dwelling for which no response had been received. The DCS found that 17.4% of the 934,564 dwellings classified as unoccupied were actually occupied and that 29.1% of the 366,527 dwellings with no responses that were classified as occupied or with occupancy status classified as unknown were actually unoccupied. Estimates for 24 subprovincial regions, based on the DCS sample, were used to adjust the occupancy status for individual dwellings using a small area synthetic estimation approach based on the CU level counts of dwellings with the appropriate occupancy status. This resulted in an increase of 3.6% in the number of occupied dwellings and a decrease of 5.2% in the number of unoccupied dwellings at the Canada level. More details on the DCS can be found in Dick (2007).

18. After this adjustment of the occupancy status by the DCS, occupied dwellings with total non-response had the number of usual residents (if not known) and all the responses to the census questions imputed by borrowing the unimputed responses from another household within the same CU including its type of questionnaire (long or short). This process is called Whole Household Imputation and it imputed 96% of the total non-response households. The other 4% of the total non-response households where no donor household was found under the Whole Household Imputation process were imputed as part of the main E&I process. Utilizing a single donor under Whole Household was more efficient computationally and was less likely to produce implausible results than using several donors as part of the main E&I process, as was done in 2001.

19. The DCS adjusted for both undercoverage (by converting unoccupied dwellings to occupied) and overcoverage (by converting occupied dwellings to unoccupied) at the dwelling level. Other studies (the Reverse Record Check (RRC) and Census Overcoverage Study(COS)) measure most sources of undercoverage and overcoverage for persons at the provincial/territorial level. The resulting RRC/COS estimates of net undercoverage are incorporated into Statistics Canada's population estimates program and are used in the allocation of billions of dollars to the provinces and territories by the federal

[2] I would like to thank Peter Dick of Statistics Canada for providing information on the DCS for this section.

government. Census publications, however, only reflect the adjustments made by the DCS since the RRC/COS results are not available soon enough.

20. Households enumerated on a long form and with total non-response to the questions asked on a sample basis (these will be called sample questions) were converted to short forms in a process called "Document Conversion" while ensuring that the percentage of long forms did not fall below 5% at the CU level. This resulted in a 0.49 % reduction in the number of long forms. Remote areas with sampling fractions somewhat less than 100% had short forms converted to long forms to achieve the expected 100% sample for these areas. This resulted in a 0.97% increase in long forms for remote areas.

21. Note that Whole Household Imputation resolved total non-response to all questions on short forms and long forms. Then Document Conversion followed by the weighting of the remaining long forms dealt with total non-response to the long form sample questions. It made sense to use weighting in the latter case since the 20% sample of long forms has to be weighted up to the population level anyway. In the 2001 Census, unoccupied dwellings were "converted" to a status of occupied by the DCS by increasing the weights for some occupied long form dwellings rather than using Whole Household Imputation. Using imputation in 2006 had the following advantages compared to weighting:

- Both short and long forms were used as donors (only long forms were used with weighting in 2001) so we were more likely to find a donor within the CU.
- With weighting in 2001, some city blocks within CUs had an increase in population while other city blocks had a decrease in population. This could have caused inconsistencies that data users could have noticed since population counts were provided at the city block level. With imputation, this did not happen.

V. SEQUENCE IN WHICH VARIABLES WERE PROCESSED IN E&I

22. Census E&I processing was carried out over a twelve month period. The census variables were partitioned into subject matter topics which were then processed sequentially or in parallel in the following order: type of dwelling, postal code, demographic variables, languages, dwelling questions, immigration, labour, mobility, aboriginal status, place of birth of parents, place of work, education, unpaid work, race, mode of transport to work and income. Each subject matter topic was processed in a series of modules. Typically a minimum change donor imputation module was preceded and followed by modules which derived variables and performed deterministic imputation (these are called pre-derive and post-derive modules). A version of the variable both before and after imputation was kept so the impact of imputation could be measured. Once variables were finalized by the modules of one subject matter topic, they could not be changed by the modules of a later subject matter topic.

23. This partitioning of E&I into subject matter topics was necessary because of

- The need for imputed responses to filter questions that determine whether other questions should be answered. Age, for example, had to be imputed before it could be determined whether responses were needed for other questions asked only of those of age 15+.
- The need to derive new variables (e.g. those related to census family structure) based on imputed responses before processing the next subject matter topic.
- The likelihood that very few households would pass the edits for all subject matter topics so that there would be few completely error-free donors.
- The prohibitive computation costs of processing all variables together.

24. This partitioning into subject matter topics, however, caused problems where the imputation of one variable has a negative impact on the imputation of other variables in later modules. For example, assigning an age of less than 15 to someone in the demographic edits occasionally caused the responses of many other questions asked only of adults to be blanked out.

VI. DEVELOPMENT OF E&I SOFTWARE

25. CANEDIT (CANadian EDITing System), based on the minimum change imputation methodology proposed by Fellegi/Holt (1976), was implemented for the 1976 Census. It performed minimum change donor imputation and was an early example of a generalized processing system where the edit rules were an input into the system rather than being hard coded into the program. This approach was taken in 1976 because of delays experienced in the 1971 Census using ad hoc approaches to imputation which required multiple runs of the editing system before all the edits would pass.

26. For the 1981 Census, SPIDER (System for Processing Instructions for Directly Entered Requirements) was introduced as a second generalized system which allowed users to specify their edits through Decision Logic Tables (DLTs). SPIDER was mainly used for deterministic imputation though it was also used to do donor imputation.

27. For the 1996 Census, CANEDIT was replaced by software called NIM (Nearest-Neighbour Imputation Methodology, see Bankier 1999) and NIM was used to carry out the minimum change donor E&I of the demographic variables. The NIM allowed, for the first time, the simultaneous hot deck imputation of qualitative and quantitative variables for large E&I problems.

28. For the 2001 Census, NIM was replaced with a generalization of the methodology called CANCEIS (Canadian Census Edit and Imputation System) which was used to perform minimum change donor imputation for labour, mobility, place of work and mode of transport variables as well as the demographic variables.

29. For the 2006 Census, CANCEIS was extended to do deterministic imputation so that it could replace SPIDER. CANCEIS was then used to process all 2006 Census E&I variables except for a few modules where custom programs performed certain specialized tasks.

30. With both NIM and CANCEIS, DLTs very similar to SPIDER DLTs were used to specify the edits. In addition, a Data Dictionary like that utilized by SPIDER was used to define labels such as 'Male' and 'Female' that could be used in the edits rather than the corresponding numeric codes for sex. Adopting an interface for specifying edits similar to the one used with SPIDER made the conversion from one system to the other much easier for the subject matter experts.

31. The extensions made to CANCEIS to replace SPIDER were done as part of an iterative, collaborative effort by the methodologists, systems analysts and subject matter experts. A version of CANCEIS would be released, used by methodology and subject matter personnel to convert SPIDER modules to CANCEIS modules and then improvements would be suggested. A new version with improvements made would be released, used again and further improvements would be proposed. This iterative approach resulted in the 2006 version of CANCEIS being developed in a timely fashion and it completely met the needs of the subject matter experts.

VII. EDIT AND IMPUTATION OF THE DEMOGRAPHIC VARIABLES[3]

32. To perform E&I of the demographic variables (relationship to person one, age, sex, marital status, common-law status), processing was carried out in three stages using

- a custom program that identified potential couples, parent/child pairs and grandparent/grandchild pairs before E&I
- CANCEIS which did minimum change donor imputation using the family structure identified by the custom program and
- the same custom program, used a second time, which identified couples, parent/child pairs and grandparent/grandchild pairs after E&I and formed them into families.

[3] Most of the results in this section are taken from the report "Analysis of the Imputation of Age (Phase 2 Pass 2) for 2006 Census Data" dated September 28, 2007 by Isabelle Michaud and Michael Banker of Statistics Canada.

33. There were also various CANCEIS pre-derive and post-derive modules which were used to perform additional clean-up of these variables or derive additional variables both before and after each of the above processes.

34. There are three types of census families:
- Couples without children
- Couples with children
- Lone Parents with children.

35. Below is a brief description of the concepts that will be respected by the demographic variables for the persons in a census family after imputation.

36. For a couple, they should be
- both adults (age $\geq$ 15) and
- both married or both common-law and
- the relationships should be appropriate for a couple (some relationships can only be married, some can only be common-law, some can be either)

In addition,
- a common-law spouse must have their partner present
- some relationships imply the married partner must be present in the household
- some relationships imply a same sex couple, some imply an opposite sex couple.

37. For a child/parent pair
- at least one parent must be at least 15 years older than the child and
- a female parent must be not more than 50 years older than a child and
- the relationships should be appropriate for a child/parent pair.

In addition,
- some relationships imply the parent should be present in the household.

38. For a grandchild/grandparent pair
- at least one grandparent must be at least 30 years older than the child and
- a female grandparent must be not more than 100 years older than a child and
- the relationships should be appropriate for a grandchild/grandparent pair.

In addition,
- some relationships imply the grandparent should be present in the household.

39. In addition, there cannot be more than a 35 age difference between siblings. Otherwise their mother would have been more than 50 years old when she gave birth to the younger sibling.

40. To resolve inconsistencies for a child/parent pair, for example, responses were borrowed from a nearest neighbour donor (i.e. a donor that resembles the failed record as closely as possible) so that
- one or more of the relationships to Person 1 were changed to ensure that there was no longer a child/parent relationship or
- either the age of the child or the age of the parent or both ages were changed.

41. Before E&I for the demographic variables was carried out, two fixes to the demographic data were performed. First, the number of same sex married couples based on the unimputed data was too large. This was because some opposite sex married couples misreported their sex. The sex was changed when there was a high probability that it was wrong based on the first name. In addition, all centenarians were manually reviewed and had their numbers reduced to be more consistent with old age security counts.

42. Family data were published based on long form results only. This was because the write-ins for the relationship question (1.8% of the population provide write-ins for less common relationships) were

only captured and coded on long forms. Because of this, the analysis which follows will focus on the long forms.

43. Figure 1 shows failure rates at the household level for the demographic edits. The edit failure rate gradually increases from around 6% with a one person household to almost 44% with an eight person household. Households that fail because of blank or invalid responses only (as represented by B/I in the legend) have a large increase in the edit failure rate between the five and six person households. This is because the paper long form only had room for five persons and the respondent was not always successful in getting a second form.

44. Figure 2 shows the non-response rate (blanks plus invalids) plus the inconsistency rates (responses imputed because they were inconsistent for two or more variables) for each demographic variable. The non-response rates are highest for marital status and common-law status while the inconsistency rates are highest for relationship and common-law status.

Figure 1: Long Form Demographic E&I Household Failure Rate

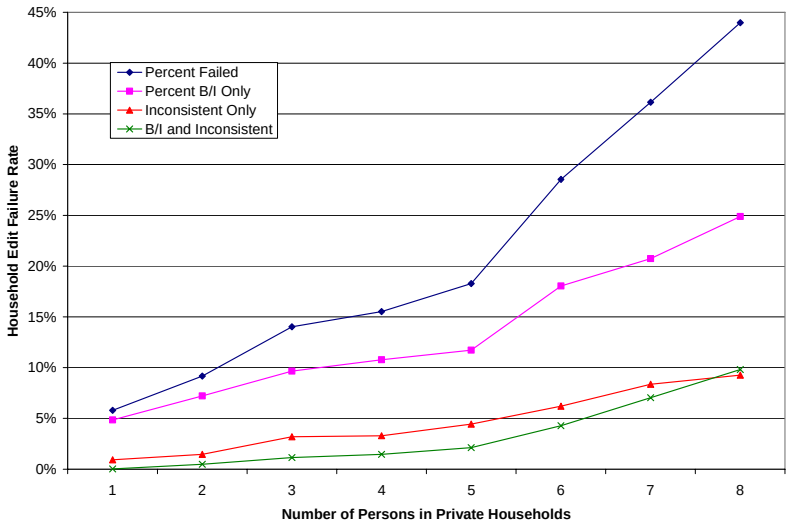
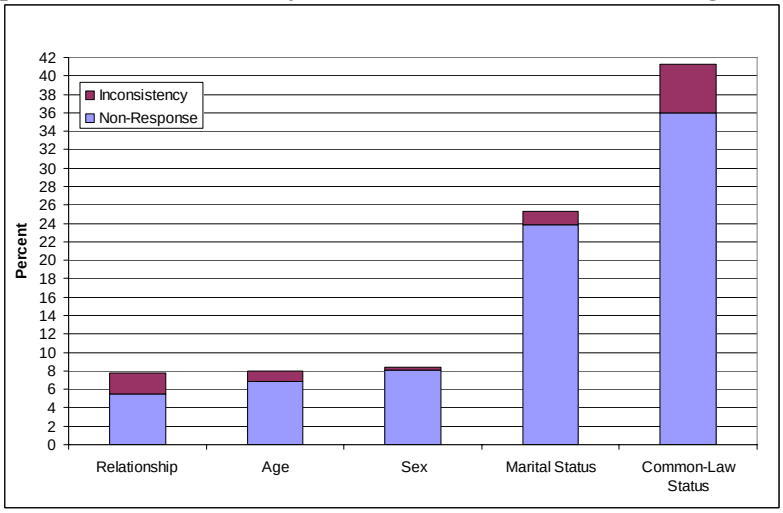


Figure 2: Non-Response and Inconsistency Rates for Persons in Failed Long Form Households



45. The effect of imputation on relationship, sex, marital status and common-law status is discussed briefly below. The impact of imputation on age is described afterwards in more detail since the patterns are particularly interesting.

46. Persons had common-law status changed from YES to NO by imputation because they said that they were also married or because they were under the age of 15. A smaller number of persons had common-law status changed from NO to YES because the other demographic variables for the household suggested that they were in a common-law relationship.

47. Persons had marital status changed from Separated or Single to Now Married because the other demographic variables for the household indicated that they were married. A smaller number of persons had marital status changed from Now Married to Single because their responses showed that they were in a common-law relationship or that they were under the age of 15.

48. A small number of persons had their sex changed to be consistent with them belonging to a same sex or an opposite sex couple.

49. Persons often had relationship changed to son/daughter from son/daughter-in-law, father/mother or father/mother-in-law because they were under the age of 15 or had some other inconsistency in their age in terms of their parents or a child. In addition, relationship was often changed to be consistent with the answers to marital status and common-law status or sex. Quite frequently, relationship was changed because Person 1 reported their relationship to the other person rather than the other person's relationship to them as they were supposed to do. Table 1, based on a real example but with age and sex modified, shows a household before and then after imputation where the reporting of relationship has been reversed. Person 1, aged 75 is followed by a father and a mother-in-law in their forties and then two grandparents, both under 15. After minimum change donor imputation, these were correctly changed into a son, daughter-in-law and two grandchildren.

Table 1: An Example Where the Reporting of Relationship has been Reversed

R2P12B	MARST	AGE	COMLAW	SEX
PERSON1	WIDOWED	75	NO	FEMALE
FATHER_MOTHER	NOW_MARRIED	45	NO	MALE
FATH_MOTH_INLAW	NOW_MARRIED	40	NO	FEMALE
GR_PRNT	SINGLE	10	NO	FEMALE
GR_PRNT	SINGLE	5	NO	MALE
PERSON1	WIDOWED	75	NO	FEMALE
*PERSON1S_SD	NOW_MARRIED	45	NO	MALE
*SON_DAUGHT_INLAW	NOW_MARRIED	40	NO	FEMALE
*GR_CHILD	SINGLE	10	NO	FEMALE
*GR_CHILD	SINGLE	5	NO	MALE

50. Next, a detailed analysis is given of the imputation of age for long form persons. Optical character recognition (OCR) was used for the first time in the 2006 Census to capture the responses from paper forms along with a small amount of manual keying for some difficult cases. A weighted sample of responses had the captured response compared to the written response for year of birth[4]. An estimated 204,260 persons (a 0.85% error rate) were in the wrong five year age range as a result of a data capture error as shown in Figure 3 below.

51. Figure 4 shows the impact of imputation on age where
- the black line is the percentage age distribution for the 99.11% of persons who did not have their age changed by imputation,
- the blue line is the percentage age distribution after imputation for the 0.61% of persons with a blank or invalid response to age before imputation,

[4] This sample was selected as part of a study by Jean-René Boudreau of Statistics Canada to assess the accuracy of the 2006 Census data capture operation.

- the green line is the percentage age distribution before imputation for the 0.28% of persons who had age imputed because of an inconsistency while
- the red line is the percentage age distribution after imputation for the same 0.28% of persons who had age imputed because of an inconsistency.

52. There are some similarities between the shape of the green line in Figure 4 and the green bars in Figure 3. But while 0.85% of the persons are estimated to have been placed in the wrong five year age range, only 0.28% of these had their age changed by imputation because of an inconsistency. Inconsistencies in the ages can only be detected by edits between parents and children, grandparents and grandchildren plus siblings or where someone provides a relationship to person 1 which implies they are an adult. There are only slight similarities between shape of the distribution of the red line in Figure 4 and the red bars in Figure 3. This is not surprising since we were only imputing a subset of the population with an error in their 5 year age range and the correct distribution of ages for this population was not known during imputation.

Figure 3: Persons in Wrong 5 Year Age Range Because of Data Capture Error

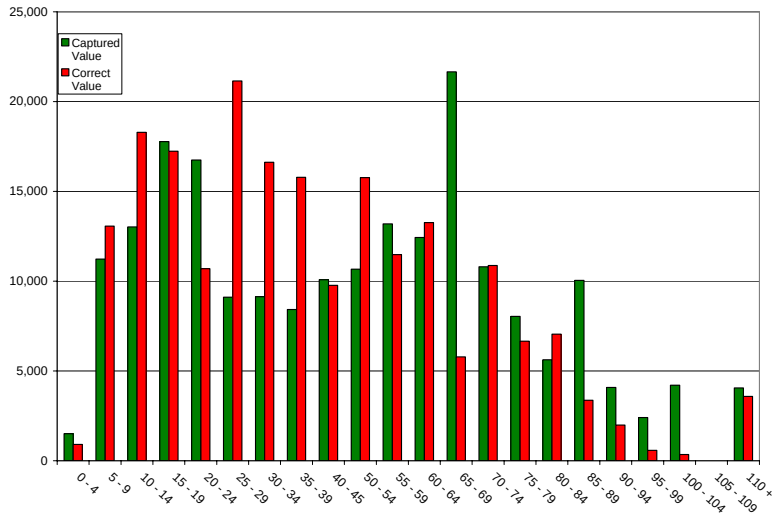
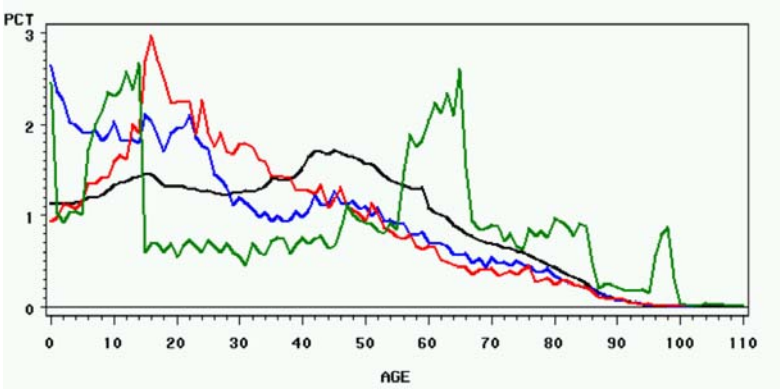


Figure 4: Impact of Imputation on AGE



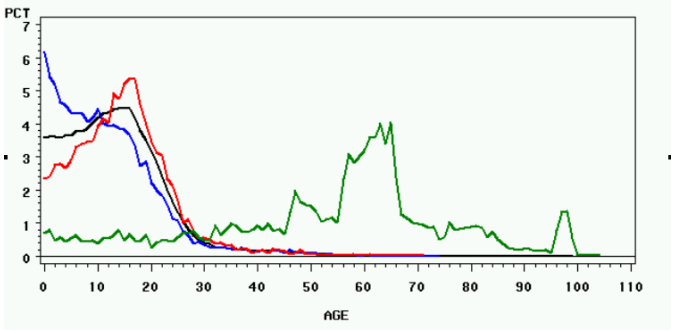
53. To get a clearer idea of what is happening with age imputation, the following three subsections will focus on the most common family statuses: HUSBAND, WIFE and TRAD_CHILD where a "traditional" child is one who has never been married and includes both children and step-children.

A. Imputation of Age for TRAD_CHILD

54. Figure 5 gives the percentage age distributions for children. Note the peaks in the green line (unimputed age for persons with ages changed because of inconsistencies) for ages of 50 or more and

how these peaks are not present in the red line (imputed ages for persons with ages changed because of inconsistencies).

Figure 5: Impact of Imputation on AGE for TRAD_CHILD

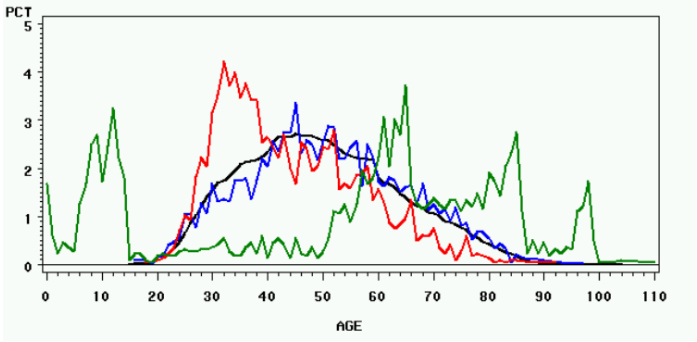


55. Random samples of households with at least one TRAD_CHILD who had their age imputed because of an inconsistency were examined. Most of these children were either older than both parents or were younger but had less than a 15 year age difference with both parents. To resolve these edit failures, a younger age had to be imputed for these children and the resulting pattern of responses seemed very reasonable in almost all of these cases. In a smaller number of cases, the child had their age changed from less than 15 to 15+ because they had a common-law status of yes. In some of these cases, it might have been better to deterministically set common-law status to no rather than increase the age. In two households, the age of the child was increased slightly so that the mother was no more than 50 years older than the child.

B. Imputation of Age for WIFE

56. Figure 6 gives the percentage age distributions for wives. A wife is required to be at least 15 years old. Note the peaks in the green line (unimputed age for persons with ages changed because of inconsistencies) for ages less than 15 and for ages older than 55. After imputation, the blue line (blanks/invalids) follows the black line (age not changed) closely as does the red line (imputed age for persons with ages changed because of inconsistencies) except for a spike around age 30.

Figure 6: Impact of Imputation on AGE for WIFE



57. Random samples of households were selected for specific negative and positive values for AGE_DIFF = AGE - AGEU which occurred with high frequency where AGEU is the age before imputation while AGE is the age after imputation. For large negative values of AGE_DIFF, frequently the age of the wife was decreased so that there was an age difference of 50 years or less with a child. The wife was also generally much older than the husband before imputation but not after. These imputation actions were very reasonable. These changes in age for these wives contributed to the large spike in the red line in Figure 6. In a few cases, however, the wife had an age which was similar to the husband before imputation but not after. In these cases, the child may have actually been a grandchild and the resulting imputation action was not that good.

58. For AGE_DIFF = -2, generally the wife had her age slightly reduced to achieve an age difference of 50 years or less between her and a child. Both the husband and wife had similar ages which might suggest that many of the children were actually grandchildren. In a few cases, the wife had her age reduced to achieve at least a 15 year age difference or at least a 30 year age difference with her parents or grandparents respectively.

59. For AGE_DIFF = 27, the wife was made older because her unimputed age was less than 15 or because there was less than a 15 year age difference between her and a child. The imputation actions in these cases were very reasonable. Again, these changes in the age of the wives contributed to the large spike in the red line in Figure 6.

60. Figures 7 and 8 plot the age of the female lone parent against the age of her child before and after imputation. Figures 9 and 10 plot the age of the wife against the age of her child before and after imputation. Dotted diagonal lines above the solid diagonal line indicate that the mother is 10, 20, 30, etc. years older than the child while dotted lines below the solid diagonal line indicate that the mother is 10, 20, 30 etc. years younger than the child. Note that after imputation, except for a few stepchildren, the female lone parent is always between 15 and 50 years older than her child as the edits require. The pattern is similar for wives except that a wife is allowed to be younger than a child if the child is a stepchild.

61. Figure 11 plots the number of children in terms of the age of the mother minus the age of the child on a logarithmic scale. These counts are before imputation, after imputation and from Vital Statistics birth records which give a lower bound on the number of children. Vital Statistics confirms that the elimination of almost all mothers with more than a 50 year age difference in census editing was the correct thing to do (only those associated with stepchildren remain). Vital Statistics also indicates that there are too many mothers with age differences in the range 45 to 50 but these are not being touched by census imputation. Census imputation actually caused an increase in the number of mothers with a 50 year age difference.

Figure 7: Age LONE_PAR_FEMALE compared to Age Child Before Imputation

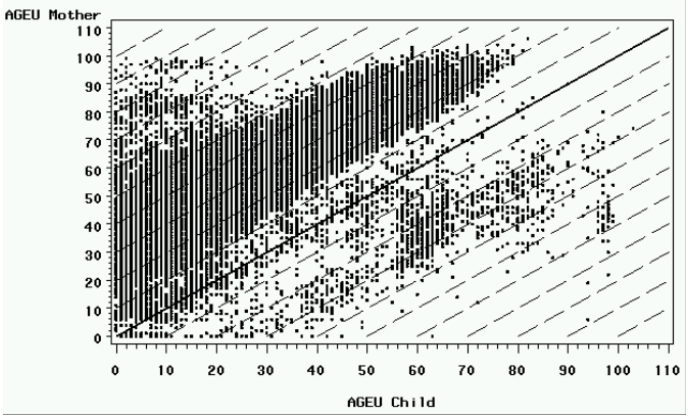


Figure 8: Age LONE_PAR_FEMALE compared to Age Child After Imputation

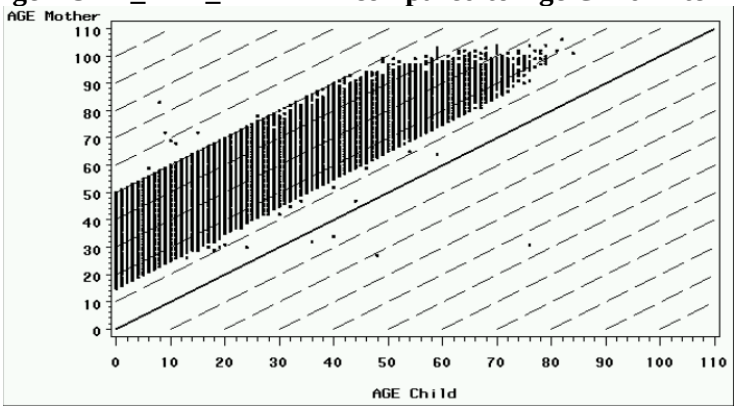


Figure 9: Age WIFE compared to Age Child Before Imputation

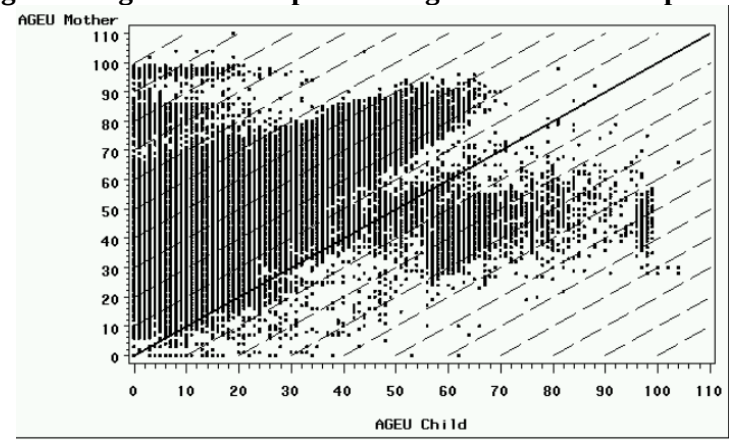


Figure 10: Age WIFE compared to Age Child After Imputation

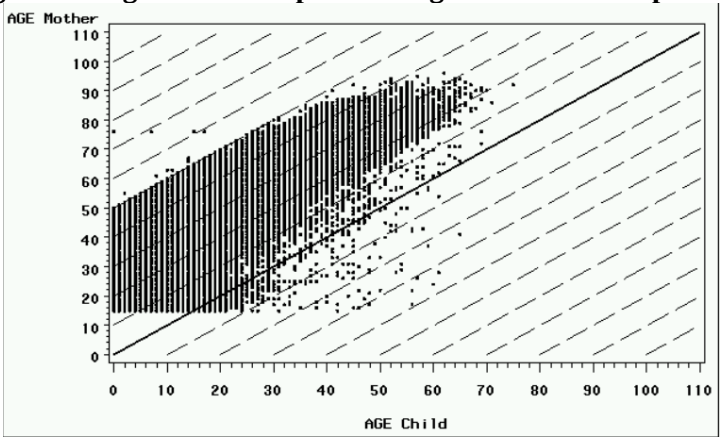
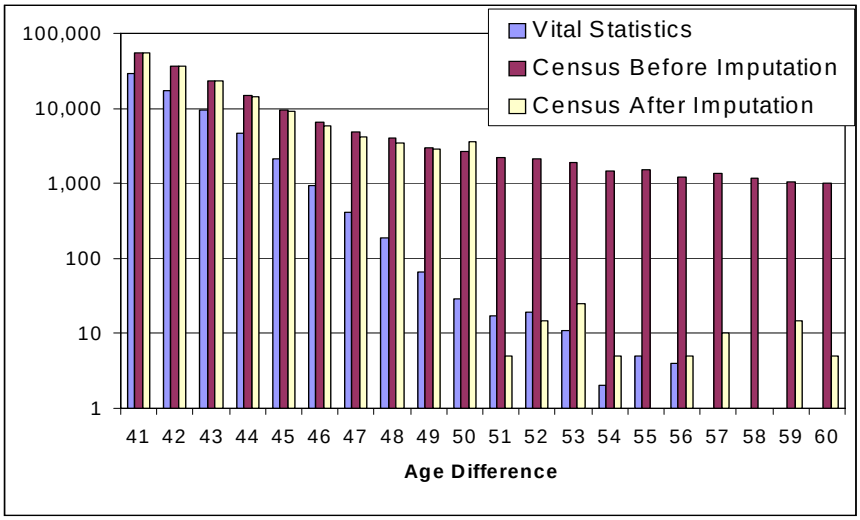


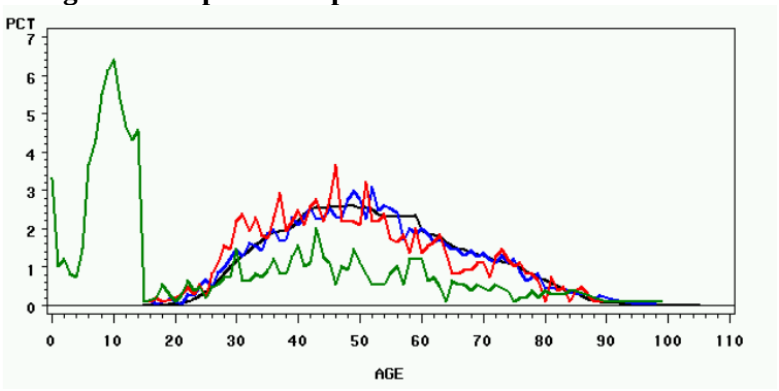
Figure 11: Number of Children in Terms of Age Difference With Mother



C. Imputation of Age for HUSBAND

62. Figure 12 gives the percentage age distributions for husbands. A husband is required to be at least 15 years old. Note the peak in the green line (unimputed age for persons with ages changed because of inconsistencies) for ages less than 15. After imputation, the blue line (blanks/invalids) and the red line (imputed age for persons with ages changed because of inconsistencies) follow the black line (age not changed) closely.

Figure12: Impact of Imputation on AGE for HUSBAND



63. Random samples of households were selected with different AGE_DIFF = AGE - AGEU ranges. For negative values of AGE_DIFF, the age of the HUSBAND was often decreased to achieve at least a 15 year age difference with one or more parents. The resulting imputation actions were reasonable. Also, the age of the HUSBAND was frequently decreased because there was more than a 35 year age difference between them and a sibling. Again, the resulting imputation actions were reasonable. For positive values of AGE_DIFF, frequently these were husbands with an age less than 15 that had this changed to an age of 15 or more. The resulting imputation actions were reasonable. It was also quite common to increase the age of the HUSBAND to achieve at least a 15 year age difference with a child or less frequently to achieve at least a 30 year age difference with a grandchild. These imputation actions were reasonable. Occasionally, the age was imputed when the reporting of the relationship had probably been reversed (see Table 1) and then the resulting imputation action was not ideal.

64. Figures 13 and 14 show the age of the husband versus the age of their child before and after imputation. It can be seen that there are some fathers who are much older than their children both before and after imputation. There was no edit for husbands like the one forbidding more than a 50 year age difference between a woman and her child because very old fathers can occur. Some of these old fathers remaining after imputation, however, are the result of data capture errors for date of birth.

65. Figures 15 and 16 compare the age of husbands to the age of wives before and after imputation. After imputation, the number of wives much older than their husbands is significantly reduced though some still remain but there is little reduction in the number of husbands much older than their wives for the reasons given in the previous paragraph.

Figure 13: Age HUSBAND compared to Age Child Before Imputation

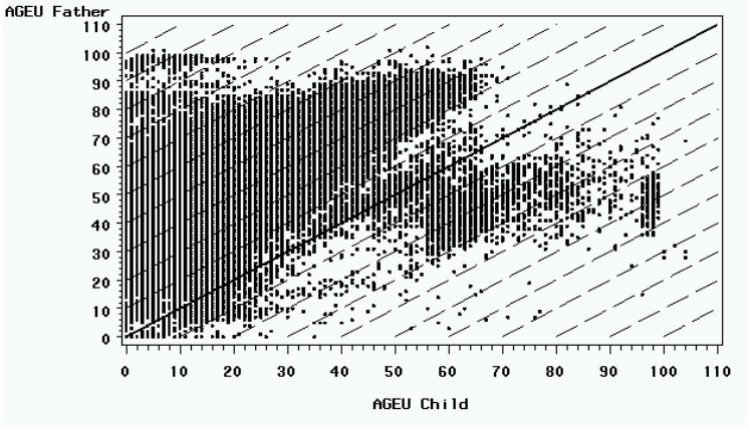


Figure 14: Age HUSBAND compared to Age Child After Imputation

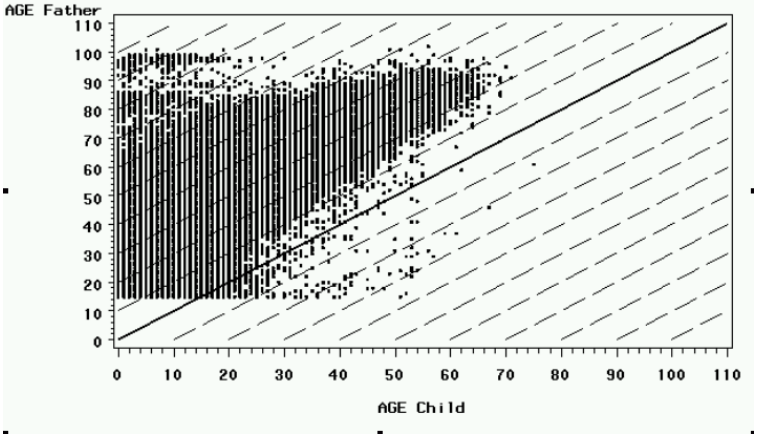


Figure 15: Age HUSBAND compared to Age WIFE Before Imputation

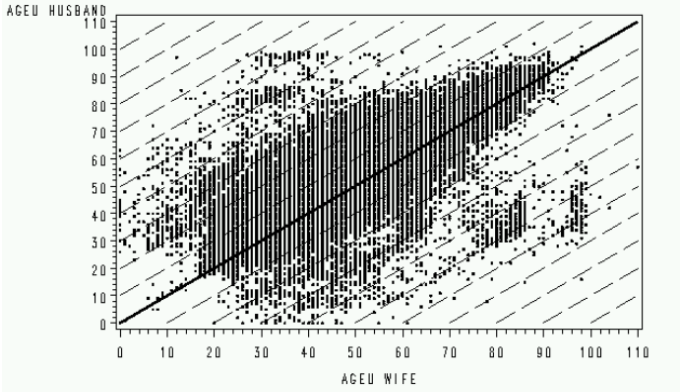
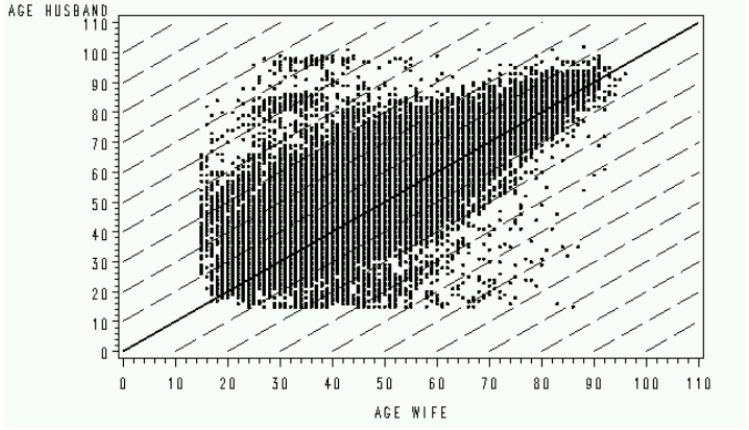


Figure 16: Age HUSBAND compared to Age WIFE After Imputation



VIII. POSSIBLE CHANGES FOR THE 2011 CENSUS

66. Some possible changes being considered for the 2011 Census are outlined below.

67. Age and date of birth may both be asked on the census form to make it easier to correct for data capture errors.

68. The 2006 and 2011 census data bases could be linked at the household level to allow targeted non-response follow-up in the field which could result in better Whole Household Imputation for total non-response households. There could also be some limited benefits when imputing for partial non-response particularly if the linkage could be done at the person level. No decision has yet been made on whether to do this.

69. An experiment, using 2006 data, is being carried out to blank out a proportion of the responses gathered as part of field edit follow-up (FEFU) and see if the original distributions can be successfully recovered by imputation where the mode of collection (e.g. internet, paper) and the date that the questionnaire was received will be used in the selection of nearest neighbours. If successful, this approach could allow less to be spent on FEFU in 2011.

70. The demographic edits and the demographic data will be studied further to determine if the edits should be made more or less stringent with an aim to maximizing the data quality.

71. Section V described why census variables are partitioned into subject matter topics and processed sequentially or in parallel. For the reasons given in Section V, however, it is not a trivial task to combine subject matter topics. If this could be done, however, it could result in better quality imputation actions and speed up processing. Two approaches will be taken to try to do this. First there will be a review of all subject matter topics to see if some can be combined and simplified. As part of this process, it will be seen if donor imputation can be used more and deterministic imputation can be used less. Second, CANCEIS may be changed so that it will retain several imputation actions for each failed record at the end of processing for one subject matter topic. These will all be processed by one or more subsequent subject matter topics (which may generate more imputation actions for each failed record) and finally only one imputation action will be randomly retained for a failed edit record at the end of the process.

72. In some complex situations, which would ideally require a manual review of all responses for a household to determine a correction, it is not possible to easily specify edits to correct a data quality problem. As an alternative, a 20% sample of the problematic cases, for example, could be manually reviewed to determine which responses are in error and should be blanked out. The other 80% of the problematic cases could have nearest neighbour imputation applied using the 20% sample as donors to determine which responses among this 80% should also be blanked out. A later donor module could then impute the blanked out responses. This technique will be tried with age data from the 2006 Census to see how well it works.

73. Currently the only records that can be used as donors are those which have passed all the edits. Sometimes, however, a record fails the edits because it is missing just one or two responses and would be quite suitable for imputing other responses. Thus failed records may be used as donors in the 2011 Census, particularly if there is a shortage of donors because of a high edit failure rate.

74. The distance measure used to identify nearest neighbour donors and then determine the best imputation actions allows different weights to be applied to the different distance functions used for each variable. A small weight indicates that a variable can be imputed without much penalty while a large weight indicates that that variable should not be imputed if possible and instead should ideally match for the failed record and the donor. These weights are fixed for all failed records within a stratum of a module. It would be desirable to allow these weights to vary by household based on the pattern of responses observed in each household. These household level weights would be assigned in a pre-derive

module. Allowing the weights to vary by household would allow, in effect, a mixture of deterministic and donor imputation to be done in the donor module.

IX. CONCLUDING REMARKS

75. The availability of sophisticated E&I programs and reduced computational costs allow subject matter experts and methodologists to do a much better job with E&I than was possible in the past. With this power comes the responsibility to make as few assumptions as possible regarding the characteristics of the non-respondents or those giving inconsistent responses and to be able to defend these assumptions. The amount of imputation done should be made clear to users as well as the sort of problems that were encountered with the data. Finally, E&I should not be viewed as a panacea such that data quality standards in the field or during processing can be lowered.

76. Even with the best efforts, an unmeasurable bias will be introduced by imputation or will remain because of uncorrected response bias or processing errors. This bias will probably be larger than the variance introduced by imputation. Thus the focus in imputation should be on the first order problem of generating the best imputation actions possible rather than the second order problem of what the imputation variance might be.

References

Bankier, M. (1999), "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses", Working Paper 24, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Italy (Rome).
(http://www.unece.org/stats/documents/1999.06.sde.htm)

Dick, P. (2007). "The 2006 Dwelling Classification Survey" Internal Report, Statistics Canada

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.
