IMPUTING NUMERIC AND QUALITATIVE VARIABLES SIMULTANEOUSLY

Michael Bankier, Jean-Marc Fillion, Manchi Luc and Christian Nadeau
Manchi Luc, 15A R.H. Coats Bldg., Statistics Canada, Ottawa K1A 0T6

KEY WORDS: Hot deck imputation, single donor, minimum change imputation, nonresponse, inconsistent response.

1. INTRODUCTION

Many minimum change hot deck imputation systems, both at Statistics Canada and internationally, are based on the imputation methodology proposed by Fellegi and Holt (1976). Examples of such edit and imputation (E&I) systems are CANEDIT (Pageau 1992) and SPIDER (Ciok 1992) used in the Canadian Census to impute qualitative variables and GEIS (Cotton 1991) used in Statistics Canada business surveys to impute numeric variables.

In preparation for the 1996 Canadian Census, the best way to carry out edit and imputation (E&I) for the basic demographic variables age, sex, marital status and relationship to person 1 was reassessed. SPIDER was designed to handle small imputation problems and could not be modified to handle E&I of the basic demographic variables. CANEDIT had been used since the 1976 Census to do E&I for these variables. While CANEDIT successfully identified and imputed the minimum number of variables, many individual imputation actions were implausible and small but important groups in the population had their numbers falsely inflated by the imputation actions. For some households (particularly those with six or more persons), CANEDIT unnecessarily used two or more donors to impute the demographic variables when only one donor was needed. This may have contributed to the implausible combinations of responses. Finally, because CANEDIT could only process qualitative variables, decade of birth had to be used in the edits. Much better edits and imputation actions would have resulted if the discrete numeric variable age could have been used in the edits.

A New minimum change hot deck Imputation Methodology (NIM) has been developed, programmed and applied on a test basis to approximately 80,000 six and eight person households from the 1991 Census. This imputation methodology takes a somewhat different approach to that used by Fellegi and Holt while at the same time capitalizing on some of their insights. The NIM will be used in the 1996 Canadian Census to carry out E&I for the basic demographic variables.

The NIM offers some significant advantages as compared to CANEDIT. It allows, given the donors available, minimum change imputation of qualitative and numeric variables simultaneously. It is less likely to falsely inflate the size of small but important groups in the population. The imputation actions for individual households are often more plausible with NIM than with CANEDIT. In addition, it can carry out minimum change imputation for larger groups of variables than CANEDIT. Finally, NIM will always perform imputation based on a single donor.

The remainder of this report compares the NIM methodology to that used by CANEDIT. Detailed comparisons were not done with GEIS (though GEIS is discussed in Section 5) since the majority of the Census variables are qualitative. Section 2 explains what the primary objectives for an imputation methodology should be. Section 3 outlines the common features of any single donor hot deck imputation methodology. Section 4 describes the NIM while Section 5 describes the CANEDIT imputation methodology. Some concluding remarks are provided in Section 6.

2. PRIMARY OBJECTIVES FOR AN IMPUTATION METHODOLOGY

Census edit rules are used to define invalid (including blank) responses for the basic demographic variables gathered for everyone in Canada. In addition, the edit rules check for responses that are inconsistent within a person and between persons in a household. A household record fails the edits if it contains invalid or inconsistent responses. Otherwise the record passes the edits. An imputation methodology is used to determine which variables to impute for each failed edit household and what values these imputed variables should take on. Usually one insists that the imputed values come from a household that passed the edits. This household will be called a donor.

Table 1 displays a household that failed the demographic edits in the 1991 Census along with the CANEDIT imputation action which is underlined. This household failed the edit rule that "The decade of birth for a son or daughter is the same or precedes the decade of birth reported for Person 1". Studying this household, the most reasonable imputation action is to change person 2's relationship to person 1 to spouse. This makes sense because person 1 and person 2 are similar in age, opposite in sex, are

married and the ages of the four daughters are reasonable. When available donors were investigated in Luc (1993), it was found that there were 97 (person 1/spouse/four child) households for every 3 (person 1/five child) households. Of the existing (person 1/five child) households, few if any would have a married daughter of age 22 present with a person 1 of age 34 and four children of ages 2 to 14. CANEDIT has thus increased the number of a rare type of household (person 1/five child) when creating a (person 1/spouse/four child household) would have been more plausible. CANEDIT, on average, will impute a four child household 1/3 of the time and a 5 child household 2/3 of the time in this situation. This is because one of three variables (person 2's relationship to person 1 and the decade of birth of person 1 or 2) can be imputed to make the household pass the edits. As described in Section 5, each of these variables has one chance out of three of being selected for imputation by CANEDIT. Thus, in this situation, CANEDIT creates implausible responses while at the same time falsely inflating the number of (person 1/five child) families.

Based on this and other similar examples, it is apparent that the objectives for an automated hot deck imputation methodology should be as follows:

(a) The imputed household should closely resemble the failed edit household. This is achieved, given the donors available, by imputing the minimum number of variables in some sense. The underlying assumption (which is not always true in practice) is that a respondent is more likely to make only one or two errors rather than several. In addition, it is important that a national statistical agency be conservative in the amount of Census data that it modifies.

(b) The imputed data for a household should come from a single donor if possible rather than two or more donors. In addition, the imputed household should closely resemble that single donor. Achieving these two objectives will tend to insure that the combination of imputed and unimputed responses for a household is plausible.

(c) Equally good imputation actions should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population. The emphasis is placed on small groups because a relatively low percentage of the demographic data is imputed in the Census. Thus even very poor imputation actions are unlikely to have much impact on large groups in the population.

These objectives are achieved under the NIM by first identifying as potential donors those passed edit households which are as similar as possible to the failed edit household. By this it is meant that the two households should match on as many of the qualitative variables as possible while having small differences between the numeric variables. (Households with these characteristics will be called close to each other or nearest neighbours.) Then, for each nearest neighbour, the smallest subsets of the non-matching variables (both numeric and qualitative) which, if imputed, allow the imputed household to pass the edits are identified. One of these possible imputation actions is randomly selected. As a result, the imputed household will be as similar as possible to the failed edit household while closely resembling the donor.

Table 1: Failed Edit Household
With 1991 CANEDIT Imputation Action Underlined

Relationship to Person 1	Sex	Marital Status	Age
Person 1	M	Married	34
Son/Daughter	F	Married	32 22
Son/Daughter	F	Single	14
Son/Daughter	F	Single	11
Son/Daughter	F	Single	6
Son/Daughter	F	Single	2

3. SINGLE DONOR HOT DECK IMPUTATION ALGORITHMS

It is useful to discuss the general features of any hot deck imputation algorithm whose aim is to impute data for a failed edit household from a single donor. Once this general algorithm is defined, alternative ways of choosing imputation actions within this common structure can be examined.

It will be assumed that the households being edited are split into a number of disjoint imputation groups that will be processed independently. For example, 2000 geographically close six person households might be placed in one imputation group.

Assume that an imputation group has F failed edit households $\underset{\sim}{V}_f$, f = 1 to F, and P passed edit households $\underset{\sim}{V}_p$, p = 1 to P. The households are classified into those which fail or pass based on J edit rules which have I variables (either qualitative or numeric) entering at least one of these J edit rules explicitly. Each failed edit household $\underset{\sim}{V}_f$ will be compared to each passed edit household $\underset{\sim}{V}_p$. For a specific $\underset{\sim}{V}_f$ and $\underset{\sim}{V}_p$, assume that I_{fp}^* of the I variables do not match. The $2^{I_{fp}^*}-1$ imputation actions possible for that $\underset{\sim}{V}_f$ and that $\underset{\sim}{V}_p$ can be listed. With $I_{fp}^* = 2$, for example, one can impute the first non-matching variable, the second non-matching variable or both non-matching variables. The possible imputation actions can be identified for each of the P passed edit households. There will then be

$$N_f = \sum_{p=1}^{P} (2^{I_{fp}^*}-1) \qquad (1)$$

possible imputation actions $\underset{\sim}{V}_{fpa}$ for a specific failed edit household $\underset{\sim}{V}_f$. A size measure will be assigned to each of the N_f possible imputation actions and one will be selected with probability proportional to these size measures. Imputation algorithms only differ in what size measure is assigned to each of the N_f possible imputation actions. It will be assumed here, however, that the size measure for imputation actions that do not pass the edits will always be set to zero.

The basic underlying assumption for any hot deck imputation algorithm is that there are donors available which closely resemble the failed edit record. It is also assumed that these donors show the correct distribution of imputed responses for the failed edit record. One of the imputation actions associated with one of these donors will be randomly selected for use with the failed edit record. If there are not enough such donors available, then donors are used which somewhat resemble the failed edit record. In extreme cases, donors are used which do not resemble the failed edit record that closely. In this situation, the required distributional information for the failed edit record is not present in the donors and it is likely that implausible imputed responses will result. Under these circumstances, no imputation algorithm will perform well.

4. DESCRIPTION OF THE NIM

The approach used by the NIM will now be more precisely described. A distance measure $D(\underset{\sim}{A},\underset{\sim}{B})$ is defined which measures the distance between the variables of the two households $\underset{\sim}{A}$ and $\underset{\sim}{B}$. With qualitative variables, the distance measure is a count of how many of the qualitative variables of $\underset{\sim}{A}$ do not equal (or match) the qualitative variables of $\underset{\sim}{B}$. With a numeric variable such as age, a value in the range 0 to 1 inclusive is added to the distance. If the age of the person in the donor household is similar to the age of the person in the failed edit household, a value close to 0 is added to the distance. Otherwise a value close to 1 is added.

The weighted average (with $0.5 < \alpha \leq 1$)

$$D(\underset{\sim}{V}_f,\underset{\sim}{V}_p,\underset{\sim}{V}_a) = \alpha D(\underset{\sim}{V}_f,\underset{\sim}{V}_a) + (1-\alpha) D(\underset{\sim}{V}_a,\underset{\sim}{V}_p) \qquad (2)$$

is calculated for each of the N_f possible imputation actions $\underset{\sim}{V}_a$ which pass the edits. A value of α equal to approximately 0.9 is chosen so that more emphasis is placed on minimizing $D(\underset{\sim}{V}_f,\underset{\sim}{V}_a)$ rather than minimizing $D(\underset{\sim}{V}_a,\underset{\sim}{V}_p)$. Those imputation actions which minimize (or nearly minimize) $D(\underset{\sim}{V}_f,\underset{\sim}{V}_p,\underset{\sim}{V}_a)$ are identified, given similar size measures and then one is randomly selected. Other imputation actions are given zero size measures and cannot be selected.

The NIM usually imputes, with a single donor, the minimum number of variables given the donors available. Often the NIM imputes the same number of variables as CANEDIT which is the theoretical minimum. Sometimes, however, CANEDIT used two or more donors to impute the minimum number of variables while the NIM was able to impute the minimum number of variables using a single donor. In a few cases, NIM imputed more than the theoretical minimum number of variables. Usually, however, this was the result of the NIM changing two ages by a little rather than one age by a greater amount so imputation actions of similar quality resulted.

The NIM ensures that the imputation action resembles both the failed edit record and the donor as closely as possible and that equally good imputation actions are selected with similar probabilities. Thus the NIM imputation actions are generally more plausible than those of CANEDIT. Also, small groups are less likely to be adversely affected.

More details on the NIM theory is provided in Bankier (1994) along with computationally efficient algorithms used to implement it.

5. DESCRIPTION OF THE CANEDIT IMPUTATION METHODOLOGY

This section describes how the Fellegi and Holt imputation methodology was implemented in CANEDIT. Certain aspects of this implementation, which are not intrinsic to the theory (and could be easily corrected), sometimes resulted in poor quality imputation actions. These are identified below. Other undesirable aspects of CANEDIT, which cannot be so easily corrected, are also discussed.

To achieve minimum change imputation, CANEDIT first analyses the edit rules to determine the theoretical minimum number of variables to impute in order for the failed edit household to pass the edits. If there is more than one minimum set of variables to achieve this, CANEDIT selects one at random and discards the others. CANEDIT searches for donors which match the failed edit household on certain variables involved in the edits that will not be imputed. It randomly selects one of the donors found in the imputation group which satisfies the matching criteria for the single minimum set of variables retained for imputation. The values from the donor household are substituted for the values in the failed edit household for the variables identified as the minimum number to impute. The matching variables are selected to ensure that the imputed household will pass the edits. This is known as primary imputation. If no donor is found which matches on these variables, CANEDIT attempts to impute the minimum set of variables sequentially using a separate donor for each variable. This is called secondary imputation. If it cannot find a suitable donor for a single variable, default imputation is used where the left-most allowable response is imputed for a variable (responses are arranged from left to right in alphabetic order).

For each variable under primary imputation that is to be imputed, auxiliary variables can be defined that the failed edit record and the donor have to match exactly. If no donor can be found that satisfies the matching criteria for a minimum change donor plus the auxiliary variables, a donor will be searched for which only satisfies the matching criteria for a minimum change donor.

Under secondary imputation, the minimum set of variables is imputed sequentially. For the first variable in the minimum set, the possible responses allowable for imputation are determined and donors with these responses are retained. Then the first retained donor encountered which matches most closely the auxiliary variables for the first variable in the minimum set is used. This process is then repeated sequentially for the other variables in the minimum set.

In summary, CANEDIT first determines which variables to impute for a failed edit household and then searches for donors. The NIM, in contrast, first searches for donors and then determines the minimum number of variables to impute given the failed edit household and the specific donors. The NIM also tries to ensure that the imputed household resembles the donor as closely as possible and that equally good imputation actions are selected with similar probabilities. It should also be noted that the NIM never resorts to secondary or default imputation. The approach used by the NIM is more data driven and is therefore less likely to create implausible imputed responses or falsely inflate the size of small but important groups in the population.

In the subsections which follow, the various components of the CANEDIT imputation methodology are analysed to determine where there are problems. The difficulties of Subsections 5.2 and 5.3 can easily be resolved. The advantages of the NIM compared to CANEDIT, however, based on the above discussion and that in Subsection 5.1, are clear.

5.1 Determining the Theoretical Minimum Number of Variables to Impute

CANEDIT can determine the theoretical minimum number of variables to impute for qualitative variables. GEIS can determine the theoretical minimum number of variables to impute for numeric variables. CANEDIT can extend its approach to discrete numeric variables by treating them as qualitative variables but it quickly becomes very expensive computationally. CANEDIT, for example, had to use decade of birth rather than age in the demographic edits for this reason. No computationally feasible technique is known that will determine the theoretical minimum number of variables to impute for a mixture of qualitative and numeric variables.

The NIM determines simultaneously the minimum number of qualitative and numeric variables to impute for a particular failed edit record and a particular donor. The problem is much simpler computationally and conceptually because if there are I_{fp}^* non-matching variables for a particular $\underset{\sim}{V}_f$ and $\underset{\sim}{V}_p$, then there are only $2^{I_{fp}^*}-1$ imputation actions that have to be considered.

It should also be noted that determining the theoretical minimum number of variables to impute without looking first at the donors means that preference will always be given to imputing one

numeric variable while in some situations imputing two numeric variables by smaller amounts may be an equally valid or better imputation action. Thus GEIS (and CANEDIT if it could handle numeric variables) will sometimes discard legitimate imputation actions.

Finally, if many variables are being imputed and there are relatively few donors, there may in fact exist no single donor which will allow the theoretical minimum number of variables to be imputed. CANEDIT will then go to secondary or default imputation. NIM will impute more than the minimum number of variables in this case but it will be from a single donor and is more likely to be a plausible imputation action.

5.2 Selecting One Minimum Set of Variables to Impute at Random Before Considering the Distribution of Responses

Both CANEDIT and GEIS randomly choose a single set of variables to impute whenever more than one such set can be found. This was done to save computational resources but is not an integral part of the theory of Fellegi and Holt. The example in Table 1 of Section 2 shows that doing this can artificially increase the size of certain small groups plus create implausible imputed responses. This, if possible, should be avoided. This can be done by considering all minimum sets of variables to impute when searching for donors. SPIDER, in fact, does this.

5.3 Searching for Donors

CANEDIT determines a subset of variables (known as matching variables) which enter the edits but will not be imputed. CANEDIT then searches for donors which match the failed edit household on all the matching variables.

This method of searching for donors is not very satisfactory. Often only a few matching variables are used. In the example of Table 1, CANEDIT only required that the donor match the failed edit household on Decade of Person 1, Relationship of Person 2 to Person 1 and Marital Status of Person 2. This is because the matching variables are chosen to ensure that the imputed household passes the edits. It does not guarantee, however, that the donors which qualify closely resemble the failed edit household. Thus CANEDIT will not necessarily select a nearest neighbour. Some of the possible damage can be mitigated by the use of auxiliary constraints but this requires the user to be aware of the problem and use the auxiliary constraints wisely.

It has also been found that CANEDIT often resorts to secondary imputation actions even when the NIM is able to impute the minimum number of variables using a single donor. This happens because CANEDIT requires that the donor match the failed edit household on all the matching variables under primary imputation and this is not always possible. With secondary imputation, however, a donor will always be found if one exists which has an acceptable value for the variable being imputed. This can result, however, in the donor matching the failed edit record on few if any variables. Also, if two or more variables are being imputed for a household, two or more donors will be used. CANEDIT used secondary or default imputation actions for 42% of the eight person households on the east regional data base while the NIM was able to impute the minimum number of variables from a single donor for 95% of the eight person households on the Ontario regional data base.

The above problems related to searching for donors could be resolved by having CANEDIT search for donors in an improved fashion (e.g. doing something similar to what the NIM does).

6. CONCLUDING REMARKS

The NIM performs minimum change hot deck imputation of qualitative and numeric data simultaneously , given the donors available, in a computationally feasible fashion. It has the potential for application to a wide range of surveys and censuses. The preliminary version of the NIM software will now be upgraded to a production system. Further study will be done to optimize parameters and the distance measures used by the NIM in preparation for its use on the demographic variables in the 1996 Canadian Census.

REFERENCES

Bankier, Mike (1994), "Imputing Numeric and Qualitative Census Variables Simultaneously", Social Survey Methods Division Report, Statistics Canada, Dated March 24, 1994.

Ciok, Rick (1992), "Spider - Census Edit and Imputation System", Social Survey Methods Division Report, Statistics Canada, Dated September 1992.

Cotton, Cathy (1991), "Functional Description of the Generalized Edit and Imputation System", Business Survey Methods Division Report, Statistics Canada, Dated July 25, 1991.

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.

Luc, Manchi (1993), "Preliminary Results on Analysing Age, Sex, Marital Status, Common-law Partner Status and Relationship to Person 1 for Some 1991 Six person Household Data", Social Survey Methods Division Report, Statistics Canada, Dated February 9, 1993.

Pageau, François (1992), "Features of the CANEDIT Software", Social Survey Methods Division Report, Statistics Canada, Dated September 1992.