ADDITIONAL DETAILS ON IMPUTING NUMERIC AND QUALITATIVE VARIABLES SIMULTANEOUSLY

Michael Bankier, Manchi Luc, Christian Nadeau and Pat Newcombe
Michael Bankier, 15Q R.H. Coats Bldg., Statistics Canada, Ottawa, Ontario K1A 0T6, Canada

KEY WORDS: Hot deck imputation, 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 and SPIDER used in the Canadian Census to impute qualitative variables and GEIS 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 a large sample of 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.

In Bankier, Fillion, Luc and Nadeau (1994), the NIM methodology is compared to that used by CANEDIT. In this paper, additional details are given on the NIM methodology. It is assumed that the reader is familiar with the 1994 paper. A technical report is available from the authors if the reader would like more information.

2. Objectives and Overview of NIM

Based on the discussion in the 1994 paper, 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. This assumption is made because 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 the imputed household is plausible.

(c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population (e.g. persons whose age is over 100).

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. (A term will be underlined when it is first defined.) 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.

More details on the NIM methodology are given in Sections 3 to 6. A brief discussion of how it can be implemented in a computationally efficient fashion is given in Section 7.

3. Variables Edited, Strata and Imputation Groups

The qualitative and discrete numeric demographic variables sex, marital status, common-law status, relationship to person 1 and age are edited for each person in the household. Both within person and between person edits are applied. A within person edit involves the variables of a single person. A between person edit involves the variables of two or more persons. Between person edits make it necessary for all demographic variables in a household to have E&I applied simultaneously. This in turn requires that the search for nearest neighbours be done at the household level rather than at the person level.

The households being edited are split into a number of disjoint strata which are further sub-divided into disjoint imputation groups that are processed independently. For example, six person households could form one stratum. This stratum is then split into imputation groups of approximately 20,000 geographically close households each (20,000 is represented by a parameter which can be changed). The donor household for a failed edit household comes from the same imputation group.

4. Specifying Edits with Decision Logic Tables

The edits can be specified either as a group of conflict rules or as a group of validity rules. Conflict rules define invalid responses (often including blanks) for individual variables plus responses that are considered inconsistent for two or more variables. If a record matches the responses given by one or more conflict rules, then it fails the edits. If it does not match any conflict rule, it passes the edits. Validity rules define combinations of responses for several variables that are considered valid and consistent. If a record matches the responses given by one or more validity rules, it passes the edits. If it does not match any validity rule, it fails the edits.

The edits for the NIM will be specified using Decision Logic Tables (DLTs). A simple example of a DLT with M = 10 propositions for rows and J = 7 edit rules for columns is given in Table 1 of Appendix A. These are, after substituting Age for Decade of Birth plus some slight simplification, the 1991 Census conflict rules for the demographic variables of a two person household. RLPER stands for relationship to person 1 while MARST stands for marital status. The number following a variable name, e.g. AGE1 or AGE2, indicates that a variable applies to Person 1 or Person 2 in the household.

A Y or N for a proposition indicates that it enters the edit rule. A Y for a proposition indicates that it is true for that edit rule. A N for a proposition indicates that it is false for that edit rule. A blank for a proposition represents "Y or N" and indicates that the proposition does not enter that edit rule. Thus propositions (6) and (7) enter Edit Rule 7 (which will be labelled as C_7) while the other propositions do not enter this edit rule. Similarly, it can be said that the variables MARST2 and AGE2 of propositions (6) and (7) enter C_7 while the other variables do not enter C_7. A variable, of course, has to enter at least one edit rule or it does not take part in the edits.

Thus, C_7 should be read as

C_7: ~(MARST2 = SINGLE) and AGE2 < 15

or as

C_7: MARST2 ≠ SINGLE and AGE2 < 15

where the two propositions that enter C_7 are connected with an "and". Also, ~ represents the negation operator which is applied if a N appears in an edit rule for a proposition. A household matches conflict rule C_7 and hence fails the edits if Person 2 is less than 15 years of age and not single.

C_1, C_4 and C_7 are within person conflict rules. C_2, C_3, C_5 and C_6 are between person conflict rules.

The 7 conflict rules in Table 1 are connected by the logical operator "or", i.e.

C_1 or C_2 or C_3 or C_4 or C_5 or C_6 or C_7

This means that a household fails the edits if it matches one or more of the seven conflict rules.

5. Distance Between Failed and Passed Edit Households

Within an imputation group, it will be assumed that F households fail the edits while P households pass the edits. The responses for the households that fail and pass the edits will be labelled by $\underset{\sim}{V}_f = [V_{fi}]$, f= 1 to F and $\underset{\sim}{V}_p = [V_{pi}]$, p = 1 to P respectively. These are I x 1 vectors containing the responses for the I variables that enter the edit rules. The distance between each failed edit record $\underset{\sim}{V}_f$ and each passed

edit record $\underset{\sim}{V}_p$ in an imputation group will be calculated for each of the F x P combinations of failed and passed edit records. Those passed edit records with the smallest distances will be considered as potential donors for the failed edit record. The weighted distance between a failed edit record and a passed edit record will be defined as

$$D(\underset{\sim}{V}_f, \underset{\sim}{V}_p) = \sum_{i=1}^{I} w_i D_i(V_{fi}, V_{pi})$$

where the weights w_i (which are non-negative) can be given smaller values for variables where it is considered less important that they match. All these weights were set to 1, however, when the NIM was tested on approximately 80,000 six and eight person households from the 1991 Census.

In the above distance measure, the distance function $D_i(V_{fi}, V_{pi})$ can be different for each variable i. In the 1996 Census, however, one distance function will be used for qualitative variables while a second distance function will be used for the numeric variables. For the qualitative variables, let $D_i(V_{fi}, V_{pi}) = 1$ if $V_{fi} \neq V_{pi}$ (the i^{th} qualitative variable does not match for the two records) and let $D_i(V_{fi}, V_{pi}) = 0$ otherwise. For the numeric age variables, $0 \leq D_i(V_{fi}, V_{pi}) \leq 1$ where $D_i(V_{fi}, V_{pi})$ will be close to or equal to 0 if the difference between V_{fi} and V_{pi} is small while $D_i(V_{fi}, V_{pi})$ will be close to or equal to 1 if the difference between V_{fi} and V_{pi} is large.

The parameterized distance measure for the age variables in the 1996 Census will now be discussed in detail. Two numeric variables will be considered as nonmatching, i.e.

$D_i(V_{fi}, V_{pi}) = 1$
if $|V_{fi} - V_{pi}| \geq maxdiff(V_{fi})$ (to be defined below)
or
if V_{fi} has a blank or invalid response or
if $V_{fi} < 15$ and $V_{pi} \geq 15$ (a child to adult conversion)
or
if $V_{fi} \geq 15$ and $V_{pi} < 15$ (an adult to child conversion).

The latter two conditions (which were added after the test on 80,000 households) discourage unnecessary child to adult conversions or adult to child conversions where a child is considered to be someone under the age of 15. Adult to child conversions result in a large amount of data being lost for questions that are only asked of adults. Child to adult conversions necessitate the imputation of responses to these questions.

If none of the above four conditions hold, then

$$D_i(V_{fi}, V_{pi}) = 1 - (1 - |V_{fi} - V_{pi}| / maxdiff(V_{fi}))^r$$

where r is a non-negative constant. If $V_{fi} > k_3$ then

$$maxdiff(V_{fi}) = k_1 + k_2 \frac{(V_{fi} - k_3)}{10}$$

while if $V_{fi} \leq k_3$ then $maxdiff(V_{fi}) = k_1$. In the test on 80,000 households, the parameters k_1, k_2 and k_3 were set to k_1 = 6, k_2 = 0 and k_3 = 30 which resulted in $maxdiff(V_{fi})$ equalling 6 for all values of V_{fi}. For the 1996 Census, consideration is being given to setting k_1 = 6, k_2 = 2 and k_3 = 30 so that the value of $maxdiff(V_{fi})$ increases as V_{fi} gets larger. Thus the matching criterion for age would be progressively relaxed as V_{fi} gets larger. This change is being considered because as a person gets older, it is less important that the person in the passed edit household match the person in the failed edit household closely on age. Also, as a person gets older, there are fewer potential donors that will have a similar age and relaxing the criterion will increase the chances of finding a suitable donor.

If r = ∞, $D_i(V_{fi}, V_{pi})$ = 0 when there is an exact numeric match and it equals 1 otherwise. Other values of r, (for example, r = 1/2, 2 or 4) allow near matches (e.g. $|V_{fi} - V_{pi}|$ = 2) to have values of $D_i(V_{fi}, V_{pi})$ close to 0. In the test on 80,000 households, r was set equal to 1/4.

To ensure the best donor households are selected, the failed edit household occupants are reordered in various ways to see which results in the smallest distance compared to a particular passed edit household. This may result in fewer pass edit households having to be examined to find potential donors. Smaller distances may result through reordering because, for example, the spouse was not listed in the correct position in the failed edit household. Or the children may be listed in ascending order based on age in one household and in descending order based on age in another household. Person 1 will not be reordered because this would require recoding the relationship to person 1 variable.

6. Selection of an Imputation Action

Each passed edit record $\underset{\sim}{V}_p$ will mismatch the failed edit record $\underset{\sim}{V}_f$ on one or more variables. Let I^* represent the number of variables which mismatch. There are $2^{I^*} - 1$ possible imputation actions. With I^* = 2 for example, one can impute the first non-matching variable, the second non-matching variable or both non-matching variables. In this section, it is discussed what

criteria should be used to select one of the imputation actions from one of the passed edit records to be the actual imputation action used.

All potential imputation actions $\underset{\sim}{V}_a = [V_{ai}]$ (which is an I x 1 vector) for a specific $\underset{\sim}{V}_f$ based on the P passed edit records $\underset{\sim}{V}_p$ will be determined. The potential imputation actions will be generated based on all possible subsets of I^* non-matching variables for each passed edit record $\underset{\sim}{V}_p$. Potential imputation actions $\underset{\sim}{V}_a$ which pass the edits will be called feasible imputation actions.

A feasible imputation action $\underset{\sim}{V}_a$ will be said to be essentially new if no subset of the variables imputed for that imputation action represents another feasible imputation action. Any feasible imputation actions which are not essentially new will be discarded and will not be considered for selection as the actual imputation action. These feasible imputation actions will be discarded because one or more variables are being unnecessarily imputed and hence the principle of making as little change to the data as possible when carrying out imputation is being violated.

It is easy to see that

$$\begin{aligned} & D(\underset{\sim}{V}_f, \underset{\sim}{V}_a) + D(\underset{\sim}{V}_a, \underset{\sim}{V}_p) \\ &= \sum_{i=1}^{I} w_i (D_i(V_{fi}, V_{ai}) + D_i(V_{ai}, V_{pi})) \\ &= D(\underset{\sim}{V}_f, \underset{\sim}{V}_p) \end{aligned}$$

Thus, as might be expected, the distance of the potential imputation action $\underset{\sim}{V}_a$ from the failed edit record $\underset{\sim}{V}_f$ plus the distance of the potential imputation action from the passed edit record $\underset{\sim}{V}_p$ equals the distance of the failed edit record from the passed edit record.

A weighted average

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

will be calculated for each potential imputation action $\underset{\sim}{V}_a$ of each passed edit record $\underset{\sim}{V}_p$ being evaluated where $0 \le \alpha \le 1$. Larger values of α will be selected if it is more important to have the minimum number of variables imputed than to have the imputation action close to a passed edit record (the latter being true helps ensure the plausibility of the imputation action). In most cases, α will be selected with a value greater than 1/2. For the test with 80,000 households, α =0.9 was used.

Let min D_{fpa} represent the minimum value of D_{fpa} when all P passed edit records $\underset{\sim}{V}_p$ and all feasible imputation actions $\underset{\sim}{V}_a$ are considered for that failed edit record $\underset{\sim}{V}_f$. Any essentially new imputation actions with D_{fpa} = min D_{fpa} will be called minimum change imputation actions.

Any essentially new imputation actions $\underset{\sim}{V}_a$ which satisfy

$$D_{fpa} \le \gamma \min D_{fpa}$$

where $\gamma \ge 1$ will be called near minimum change imputation actions. For the test with 80,000 households, γ was set equal to 1.1. Values of γ greater than 1 are allowed because the near minimum change imputation actions, for practical purposes (particularly with numeric variables), are nearly as good as the minimum change imputation actions. Imputation actions which are not near minimum change imputation actions are discarded because the principle of making as little change to the data as possible when carrying out imputation is being violated.

A size measure

$$R_{fpa} = (\min D_{fpa}/D_{fpa})^t$$

will be defined for each of the near minimum change imputation actions generated by the P passed edit records available in the imputation group. We will select a single near minimum change imputation action $\underset{\sim}{V}_a$ for that failed edit record $\underset{\sim}{V}_f$ with probability proportional to R_{fpa} . If t = 0, all near minimum change imputation actions $\underset{\sim}{V}_a$ are selected with equal probability. If t = ∞, then all minimum change imputation actions are selected with equal probability and all other imputation actions $\underset{\sim}{V}_a$ where D_{fpa} > min D_{fpa} are selected with zero probability. A value of t somewhere between these two extremes will usually be chosen so that minimum change imputation actions will be selected with somewhat higher probability than imputation actions with D_{fpa} close but not equal to min D_{fpa}. In the test of 80,000 households, t was set equal to 1.

7. Implementing the NIM Efficiently

In a technical report available from the authors, it is shown how to implement the NIM efficiently. Some of the techniques used are described below. The test on 80,000 households demonstrated the effectiveness of these techniques.

In practice, it is too costly to evaluate, for each failed edit record, the imputation actions of all passed edit records. Often a sufficient number of nearest neighbours are discovered by examining just the 1000 passed edit households geographically closest to the failed edit household. Also, usually only the imputation actions for the closest nearest neighbours (in terms of

the distance measure) have to be assessed because only they will generate near minimum change imputation actions.

Before evaluating the imputation actions for a nearest neighbour and a failed edit record, it is possible to drop many edit rules that none of the potential imputation actions for that pair will match. In addition, the remaining edit rules may sometimes be further simplified by dropping propositions. The algorithm to drop and simplify edit rules in a DLT will be illustrated by continuing with the example in Appendix A. Table 2 provides a household which fails the Table 1 edits (i.e. it matches Edit Rule 5) and a household which is the nearest neighbour of the failed edit household. The nearest neighbour matches the failed edit household exactly on marital status for the two persons and nearly matches on the ages of the two persons. Because there are 5 non-matching variables, there are $2^5 = 32$ potential imputation actions. We wish to determine which of these 32 imputation actions pass the edits of Table 1 and result in the smallest number of variables possible being imputed.

To do this, the Table 1 edits will be simplified by dropping any edit rules that none of the 32 possible imputation actions can match. To do this, the propositions will be assessed consecutively. If a proposition is true for all possible imputation actions, any edit rules with a N for that proposition can be dropped because none of the imputation actions will match those edit rules. The proposition can also be dropped because the remaining edit rules match that proposition for all possible imputation actions. Alternatively, if a proposition is false for all possible imputation actions, any edit rules with a Y for that proposition can be dropped because none of the imputation actions will match those edit rules. The proposition can also be dropped because the remaining edit rules match that proposition. Additional propositions can be dropped if they do not enter any of the remaining edit rules. This algorithm will now be illustrated by points (a) to (f).

(a) Proposition (1) in Table 1 is true for the nearest neighbour and false for the failed edit household. Proposition (2) is true for the failed edit household and false for the nearest neighbour. Hence no rules or propositions can be discarded based on Propositions (1) and (2). Proposition (3) however, is false for both the failed edit household and the nearest neighbour. Thus Edit Rule 6 can be dropped because it has a Y for Proposition (3). In addition, Proposition (9), which does not enter the remaining edits, and Proposition (3) can be discarded to generate Table 1a in Appendix A.

(b) Next, it is known that Propositions (4) and (5) are not true for both the failed edit household and the nearest neighbour. Thus these two propositions can be crossed out of Table 1a along with Edit Rules 1 and 2 (which each have Y's for one of these propositions).

(c) Proposition (6) is false for both the failed edit household and the nearest neighbour. No edit rule, however, has a Y entry. Thus no edit rules can be dropped based on Proposition (6). Proposition (7) is false for both the failed edit household and the nearest neighbour. Therefore Edit Rules 4 and 7 and Propositions (6) and (7) can be crossed out of Table 1a.

(d) To assess Proposition (8), the four possible imputation actions for AGE1 and AGE2 have to be considered.

Neither AGE1 or AGE2 imputed 34 - 32 = 2
AGE1 not imputed but AGE2 imputed 34 - 33 = 1
AGE1 imputed but AGE2 not imputed 37 - 32 = 5
AGE1 and AGE2 imputed 37 - 33 = 4

It can be seen that the age difference is less than 15 for all four possible imputation actions. Thus Proposition (8) is always satisfied regardless of the imputation action. No edit rules can be dropped because of Proposition (8). Rule 5, however, is simplified when Proposition (8) is crossed out.

(e) Proposition (10) is sometimes true and sometimes it is false depending on the imputation actions for SEX1 and SEX2. Thus no additional propositions or edit rules can be dropped based on Proposition (10).

(f) Thus only Edit Rules 3 and 5 and only the variables RLPER2, SEX1 and SEX2 have to be considered when assessing which imputation actions pass the edits. Thus, a maximum of $2^3 = 8$ imputation actions involving the 3 variables remaining in Table 1a (rather than 32 imputation actions) have to be considered.

Usually only a small subset of the potential imputation actions based on the simplified DLTs have to be assessed. Most imputation actions can be dropped because they are not essentially new or are not near minimum change imputation actions.

REFERENCES

Bankier, M., Fillion, J.-M., Luc, M. and Nadeau, C. (1994), "Imputing Numeric and Qualitative Variables Simultaneously", Proceedings of the Section on Survey Research Methods, American Statistical Association, 242-247.

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.

Appendix A:

Table 1: Two Person Demographic Variables Conflict Rules

Propositions	Edit Rules 1	2	3	4	5	6	7
(1) RLPER2=SPOUSE	Y	Y	Y	Y			
(2) RLPER2=CHILD					Y		
(3) RLPER2=PARENT						Y	
(4) MARST1≠NOW MARRIED		Y					
(5) MARST2≠NOW MARRIED	Y						
(6) MARST2=SINGLE							N
(7) AGE2<15				Y			Y
(8) AGE1-AGE2<15					Y		
(9) AGE2-AGE1<15						Y	
(10) SEX1=SEX2			Y				

Table 2
Failed Edit Household - Fails Edit Rule 5 Above

Relationship to Person 1	Sex	Marital Status	Age
Person 1	M	NOW MARRIED	34
CHILD	F	NOW MARRIED	32
Nearest Neighbour - Variables That Do Not Match the Failed Edit Household Are Underlined			
Person 1	F	NOW MARRIED	37
SPOUSE	M	NOW MARRIED	33

Table 1a: Simplified Two Person Demographic Variables Conflict Rules

Propositions	Edit Rules 1	2	3	4	5	7
(1) RLPER2=SPOUSE	~~Y~~	~~Y~~	**Y**	~~Y~~		
(2) RLPER2=CHILD					**Y**	
~~(4) MARST1≠NOW MARRIED~~		~~Y~~				
~~(5) MARST2≠NOW MARRIED~~	~~Y~~					
~~(6) MARST2=SINGLE~~						~~N~~
~~(7) AGE2<15~~				~~Y~~		~~Y~~
~~(8) AGE1-AGE2<15~~					~~Y~~	
(10) SEX1=SEX2			**Y**			