

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

EXPERIENCE WITH THE NEW IMPUTATION METHODOLOGY USED IN THE 1996
CANADIAN CENSUS WITH EXTENSIONS FOR FUTURE CENSUSES

Submitted by Statistics Canada¹

Contributed paper

I. INTRODUCTION

1. Many minimum change donor imputation systems are based on the imputation methodology proposed by Fellegi and Holt (1976). For example, CANEDIT and GEIS at Statistics Canada and DISCRETE and SPEER at USBC are based on the Fellegi/Holt imputation methodology. A New Imputation Methodology (NIM) was used in the 1996 Canadian Census to carry out Edit and Imputation (E&I) for the variables age, sex, marital status, common-law status and relationship. These demographic variables were successfully processed over a one-month period for eleven million households. NIM allowed, for the first time, the simultaneous hot deck imputation of qualitative and numeric variables for large E&I problems.

2. A typical E&I problem is displayed in Table 1 for a 6 person failed edit household (only the first three people are displayed). In Table 1, there is a blank response for marital status, and the age of the mother is inconsistent with the age of her son (Person 1). Data borrowed from a household that passed the edits (which will be called a donor), is used to impute (see Table 2) a marital status of widowed for the mother plus increase her age to 59. (A term will be underlined when it is first defined.) Various subsets of the variables are imputed to determine which is the optimum imputation for a failed edit household. Each of these subsets, when imputed, will be called an imputation action.

Table 1:Failed Edit Household

<u>Relationship</u>	<u>Marital Status</u>	<u>Age</u>
Person 1	Married	<u>38</u>
Spouse	Married	35
<u>Mother</u>	<u>Blank</u>	<u>41</u>

Table 2:Imputed Household

<u>Relationship</u>	<u>Marital Status</u>	<u>Age</u>
Person 1	Married	38
Spouse	Married	35
Mother	<u>Widowed</u>	<u>59</u>

3. The Fellegi/Holt algorithm first determines the minimum number of variables to impute and then performs the imputation, possibly by searching for donors. NIM, in contrast, first searches for donors and then determines the minimum number of variables to impute given the donors. Changing the order of these

¹ Prepared by Michael Bankier.

operations allows NIM to solve larger and more complex E&I problems. NIM does require donors, however, to be able to carry out imputation.

4. In this paper, the relatively simple algorithms used to implement NIM in a computationally efficient way will be illustrated using the above example. Section II gives the objectives and an overview of NIM. Section III illustrates NIM with this example. Section IV compares NIM to several implementations of the Fellegi/Holt methodology. Section V discusses the identification of couples before editing. Section VI outlines the performance of NIM during the 1996 Census. Section VII describes a new NIM prototype that was written to allow us to test new features for the 2001 Census. Section VIII outlines how NIM could be extended to carry out imputation for edits with a large number of numeric variables. Section IX provides some concluding remarks. Additional details on the NIM methodology are given in Bankier et al (1994, 1995, 1996 and 1997). A technical report is available from the author if the reader would like more information.

II. OBJECTIVES AND OVERVIEW OF NIM

5. 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.

(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 objectives will tend to ensure 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).

6. These objectives are achieved under 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. For a specific failed edit household/nearest neighbour pair, the only candidates for imputation are, of course, those variables that do not match. Then, for each nearest neighbour within a geographical region, 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. An imputation action that passes the edits will be called feasible. One of these feasible imputation actions which imputes the smallest number (or near the smallest as defined by the distance D_{fpa} given in Section III) of variables possible (which will be called a near minimum change imputation action or NMCIA) 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.

7. These near minimum change imputation actions can be identified efficiently for each nearest neighbour being considered as a donor for the failed edit household as follows (see Section III for more details):

(a) Only edit rules that one of the possible imputation actions can fail are retained for each failed household/nearest neighbour pair. This results in many fewer edit rules being needed to evaluate the imputation actions.

(b) Variables most likely to need imputation are considered first. Thus, blanks and invalids are imputed first followed by variables which enter the edits that the household failed (since at least one of these has to be imputed) and finally the other variables.

(c) When generating imputation actions for a failed edit household/nearest neighbour pair, only those which are:

- near the optimum (i.e. are near minimum change)
- and are essentially new (i.e. no subset of the variables being imputed would pass the edits)

are evaluated for feasibility. Imputation actions that are not essentially new are discarded because one or more variables are being unnecessarily imputed. This violates the principle of making as little change to the data as possible.

III. AN EXAMPLE ILLUSTRATING THE NIM ALGORITHM

8. The failed edit household displayed in Table 1 will be used to illustrate the NIM algorithm. Edits are specified using decision logic tables (DLTs) as illustrated by Table 3. A DLT can be described as a matrix where the first column is a list of propositions (such as $Relat(3) = Mother$) followed by columns of Y's, N's and dashes that each represent an edit rule. The Table 1 household matches and hence fails the leftmost edit rule in Table 3, i.e. Person 3 is the mother of Person 1 ($Relat(3) = Mother$) but the age difference between the mother and Person 1 is less than 15 years ($Age(3) - Age(1) < 15$). This is called a between person edit rule because the responses of two persons are compared. An edit rule which compares the responses of a single person (the second edit rule in Table 3, for example) will be called a within person edit rule. An edit rule which, if matched, causes the household to fail/pass, will be called a conflict rule/validity rule.

9. A search among the passed edit households is done to identify the nearest neighbours to the Table 1 household. Preference is given to those households which are geographically close. One of these nearest neighbours is listed in Table 4 above. The five responses in Table 4 that do not match the responses of Table 1 (and hence are the only candidates for imputation) are underlined. The distance between the failed edit

Table 3: Decision Logic Table of Edit Rules

Relat(3) = Mother	Y	Y	-	-
Age(3) - Age(1) < 15	Y	-	-	-
Age(3) < 30	-	Y	-	-
Relat(3) = Grandmother	-	-	Y	Y
Age(3) - Age(1) < 30	-	-	Y	-
Age(3) < 45	-	-	-	Y

Table 4: Nearest Neighbour to Table 1 Household

<u>Relationship</u>	<u>Marital Status</u>	<u>Age</u>
Person 1	Married	<u>36</u>
Spouse	Married	<u>37</u>
<u>Mother-In-Law</u>	<u>Widowed</u>	<u>59</u>

household and this nearest neighbour (which is a measure of the number of non-matching variables) is $3 + 0.1 + 0.1 = 3.2$. The two 0.1 terms are for the two ages that differ by 2 years (and hence are near matches) while the count of three is for the other three variables that do not match.

10. Using the nearest neighbour, the blank response is imputed. The resulting imputation action still fails the first edit of Table 3. If it had passed the edits, we would have stopped since any other imputation would have not been essentially new in terms of this imputation action.

11. To make the evaluation of other imputation actions more efficient, any edit rule that no imputation action will fail (based on the Table 1 failed household and the Table 4 nearest neighbour) will be dropped. If a proposition is always true for all possible imputation actions, any edit rule with a N for that proposition can be immediately discarded, along with the proposition. Similarly, if a proposition is always false for all possible imputation actions, any edit rule with a Y for that proposition can be immediately discarded, along with the proposition. For example, person 3 is age 41 in Table 1 and age 59 in Table 4. Hence the third proposition ($Age(3) < 30$) will never be true and hence the second edit rule can be discarded as can the third proposition. Similarly, neither Table 1 or Table 4 have any grandmothers present and thus the third and fourth edit rules of Table 3 can be discarded as can propositions 4 to 6 which do not enter any of the

remaining edit rules. Thus the only Table 3 edit rule and propositions remaining are the rule originally failed by the Table 1 household along with the first two propositions which are sometimes true and sometimes false depending on the imputation action. When this process is repeated with all six person household edits (240 edits in 62 DLTs similar to the ones used in the 1991 Census), only two edit rules (see Table 5) and four propositions remain. The first edit rule in Table 5 is the simplified edit rule retained from Table 3. The second simplified edit rule in Table 5 comes from one of the other 62 DLTs. This edit rule originally had a proposition indicating that the second person had to be the spouse of Person 1. This proposition was dropped because the second person is the spouse of Person 1 in both the Table 1 and 4 households.

Table 5: Edit Rules Remaining After Simplification

Relat(3) = Mother	Y -
Age(3) - Age(1) < 15	Y -
Relat(3) = Mother-in-Law	- Y
Age(3) - Age(2) < 15	- Y

12. The $2^4 - 1 = 15$ possible imputation actions based on the four variables (Relat(3), Age(1), Age(2) and Age(3)) which enter the two edits of Table 5 will be evaluated. The 7 imputation actions based on the three variables (Age(1), Age(3) and Relat(3)) which enter the first edit rule of Table 5 (which is the edit rule that the Table 1 household originally failed) will be evaluated initially. Age(1) will be imputed first (since the age changes by only 2 years) followed by Age(3) and then Relat(3). Thus, if (0, 0, 1), (0, 1, 0) and (1, 0, 0) represent respectively the imputation of Age(1), Age(3) or Relat(3) alone, the 7 possible imputation actions and their order of imputation will be (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0) and (1, 1, 1). Only some of these imputation actions are generated in practice as will now be shown. After evaluating the first two imputation actions, it is found that (0, 0, 1) fails the edits while (0, 1, 0) passes the edits. As a result, the imputation actions (0, 1, 1), (1, 1, 0) and (1, 1, 1) are not generated or evaluated because they would not be essentially new. The two remaining imputation actions, (1, 0, 0), (1, 0, 1), are then found to not pass the second edit of Table 5. Because Age(2) enters the second edit of Table 5, it is now considered for imputation. With the first three variables, four imputation actions out of eight were retained, with (0, 1, 0) passing the edits while (0, 0, 0), (1, 0, 0), (1, 0, 1) failed the edits. Thus, the three imputation actions (1, 0, 0, 0), (1, 1, 0, 0), (1, 1, 0, 1) will be generated from these three failing imputation actions where the leftmost 1 represents the imputation of Age(2). It is found that none of these pass the edits and hence the process stops with (0, 0, 1, 0) being the only imputation which passes the edits. In practice, two other checks are done to drop additional imputation actions. If an imputation action is not a NMCIA, it is dropped before applying the edits and no additional imputation actions are generated from it. Also, if an imputation action fails an edit and all the variables in that edit have already been considered for imputation, the imputation action will be dropped (along with the edit rule) because the imputation of additional variables will not allow the resulting imputation action to pass that edit. This process of dropping edit rules, propositions (and hence variables) first and then generating and evaluating only a subset of the imputation actions, results in a very efficient minimum change imputation algorithm for large problems (as is indicated in Table 6 below).

13. The process of identifying imputation actions is repeated with a number of other nearest neighbour households. Let D_{fa} represent the distance from the imputation action to the failed edit household (i.e. a measure of how many variables are imputed). Let D_{ap} represent the distance of the imputation action to the nearest neighbour used (i.e. a measure of plausibility). The n (with $n = 5$ in 1996) imputation actions with the smallest D_{fpa} are retained where

$$D_{fpa} = \alpha D_{fa} + (1 - \alpha) D_{ap}$$

The parameter α (which can fall in the range (0.5, 1]) was set to 0.9 in the 1996 Census to place more importance on imputing the minimum number of variables. Then one of these n imputation actions is randomly selected to be the actual imputation action used for the failed edit household.

14. The computational efficiency of the NIM algorithm as the household size increases is illustrated by Table 6. These CPU times are standardized in terms of the time taken to process a one person household. Thus, to perform E&I on a three person household is 2.3 times more costly than performing E&I on a one-

person household. The number of edit rules increases rapidly as the household size increases because between person edit rules have to be generated for all possible pairs of persons in a household (this will be illustrated for Table 8 in Section 5). Thus there are 307 edit rules for a three person household and 2435 edit rules for a six-person household. While there are 8 times the number of edit rules for a six person household compared to a three person household, the computational costs increase by only a factor of 5. The computational costs for seven and eight person households are similar to six person households because of a shortage of donors for large households.

Table 6: NIM cost as Household Size Increases

<u>Household Size</u>	<u>Number of Edit Rules</u>	<u>Standardized Time in terms of Time for 1-Person Hhld</u>
1	9	100
2	49	129
3	307	230
4	787	459
5	1494	566
6	2435	1,005
7	3616	1,182
8	5043	941

IV. COMPARISON OF NIM AND FELLEGI/HOLT IMPLEMENTATIONS

15. In previous Censuses, CANEDIT, an implementation of the Fellegi/Holt algorithm, was used to carry out E&I for the demographic variables. NIM and CANEDIT imputation actions were compared for 12,000 failed edit households. Approximately 98% had the same number of variables imputed. The majority of the remaining variables had one additional variable imputed by NIM because of the more rigorous NIM edits based on age rather than decade. (CANEDIT used decade rather than age in the edits because the computational costs were otherwise too large.)

16. In a few cases, NIM will impute more than the minimum number of variables if this results in a more plausible imputation action. This is illustrated in Table 7 below. The household fails the edit that states that there should be at least a 15-year age difference between the parent and the child. The CANEDIT imputation increased the age of Person 1 from 35 to 45 by changing the decade of birth. This results in the CANEDIT edit being satisfied that the parent should be born in an earlier decade than the child. NIM changed person 3 to the wife of Person 1 plus the marital status of the couple is changed. This created a more plausible imputation action than CANEDIT.

Table 7: Imputing More Than the Minimum

<u>Failed Edit Household</u>			<u>CANEDIT Imputation</u>			<u>NIM Imputation</u>		
<u>Relationship</u>	<u>Marital Status</u>	<u>Age</u>	<u>Relationship</u>	<u>Marital Status</u>	<u>Age</u>	<u>Relationship</u>	<u>Marital Status</u>	<u>Age</u>
Person 1	Divorced	<u>35</u>	Person 1	Divorced	<u>45</u>	Person 1	<u>Married</u>	35
Son	Single	8	Son	Single	8	Son	Single	8
<u>Daughter</u>	Widowed	<u>36</u>	Daughter	Widowed	36	<u>P1's Spouse</u>	<u>Married</u>	36

17. The advantages of NIM can be summarized as follows. Due to its efficiency, its costs increase approximately linearly to the number of edit rules. With the implementations of the Fellegi/Holt methodology, the costs increase exponentially. With NIM, relatively simple algorithms are used while sophisticated linear programming techniques are required to implement the Fellegi/Holt methodology. Fellegi/Holt always imputes the minimum number of variables. NIM will occasionally impute more than the minimum if this results in a more plausible imputation action. NIM can be extended easily (as will be shown in Section VIII) to E&I problems involving solely numeric variables or E&I problems involving a large number of both numeric and qualitative variables. The Fellegi/Holt methodology is not easily extended because of computational considerations.

V. IDENTIFYING COUPLES BEFORE NIM PROCESSING

18. In 1996, a program, which ran prior to NIM, identified potential non-unique couples (e.g. there can be more than one daughter-in-law and son couple in a household) by assigning a score to all pairs of persons in a household. If the relationships, marital statuses, common-law statuses, sexes and ages indicated that a pair was likely a couple, a large score was assigned. The potential couples with the largest scores were assigned identical values for a new unimputable **Couple** variable. Then, conflict rules, similar to those in Table 8 were applied to determine if the potential couple should be retained. The application of these edit rules resulted in either the couple being retained with appropriate characteristics (i.e. opposite sex, both married or both common-law) or the couple was discarded by changing, for example, the Son/Daughter-In-Law to a Son/Daughter or perhaps a Lodger. In certain cases, potential couples with large scores had the relationship of one person blanked out before E&I if it was inappropriate in terms of the relationship of the other person. This gave a greater opportunity for an appropriate relationship to be imputed.

Table 8: Between Person Edit Rules for “Son/Daughter - Son/Daughter-In-Law” Potential Couples

Couple(#1)=Couple(#2)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Relat(#1)=S/D	Y	Y	Y	Y	Y	Y	Y	Y	N
Relat(#2)=S/D-In-Law	Y	Y	Y	Y	Y	Y	Y	N	Y
Sex(#1) = Sex(#2)	Y	-	-	-	-	-	-	-	-
Marital Status(#1) = Married	-	Y	N	-	-	N	-	-	-
Marital Status(#2) = Married	-	N	Y	-	-	-	N	-	-
Common-Law Status(#1) = Yes	-	-	-	Y	N	N	-	Y	-
Common-Law Status(#2) = Yes	-	-	-	N	Y	-	N	-	Y

19. Edits were specified in the generic form given in Table 8, but then a program prior to NIM exploded the generic Decision Logic Tables (DLTs) into a number of replicates, one for each possible pair of persons in the household. With Table 8 and a 6 person household for example, the variables (#1,#2) were replaced with (2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), (5,6). In addition, an equal number of permutations, i.e. (3, 2), (4,2) etc. were generated for a total of 20 DLTs with #1 = 5 and #2 = 6 indicating, for example, that the edits of Table 8 were to be applied to the persons in the 5th and 6th positions in the household. The necessity to replicate the generic DLTs resulted in the large number of edit rules listed in Table 6 for the larger households. Couples were identified prior to E&I to allow the most likely couples to be targeted during E&I. Doing this also reduced the number of edit rules which would have otherwise been much larger than the number listed in Table 6.

VI. PERFORMANCE IN THE 1996 CENSUS AND OBJECTIVES FOR FUTURE CENSUSES

20. NIM successfully edited and imputed eleven million households during a one-month period early in 1997. For the 850,000 private households in the Atlantic provinces, for example, 1%, 10% and 2% failed for total non-response, partial non-response only, and inconsistent responses respectively. The processing went very smoothly thanks to a very competent implementation of the methodology by Sylvie Rivest, the systems analyst, and as a result of the hard work by the methodologists and subject matter specialists in developing the edit rules and then exhaustively testing them on approximately one million households from the 1991 Census. The subject matter specialists and the methodologists were happy with the quality of the imputation actions.

21. Based on this successful experience, it has been decided for the 2001 Canadian Census to generalize the NIM software to process a wider range of variables (place of work, mode of transport and labour force) which proved problematic with the SPIDER E&I system in 1996. SPIDER, loosely based on the Fellegi/Holt methodology, was used to edit all the 1996 Census variables and imputed all of them with the exception of the demographic variables which used NIM. Because of the relatively low number of edits that SPIDER could apply simultaneously, it was necessary to split the 2435 edits for six persons households into six parts and process them separately. The NIM software, in contrast, was able to apply all the edits simultaneously when carrying out imputation. For this reason, NIM will be extended for 2001 to allow it to perform the editing without using SPIDER. For the 2006 Census, it is planned to use NIM to process all Census variables.

VII. NIM PROTOTYPE FOR THE 2001 CENSUS

22. In the summer of 1997, a new NIM prototype was completed which performed E&I without using SPIDER for the editing. It is written completely in the C programming language, runs off flat files rather than a proprietary data base and, as a result, is portable with few changes across platforms. Two other features introduced to the prototype are noteworthy. Both the editing and imputation are carried out directly from the generic DLTs without exploding them into replicates for each possible pair of persons. This offers significant computational efficiencies. In addition, the concept of essential to impute variables was introduced. After imputing the blanks and invalids present in a household, each failing edit rule is analysed to determine if there is only a single variable available which will allow that edit rule to pass (usually because the other variables in the edit rule have already been imputed or have identical values for the nearest neighbour and failed edit household). These essential to impute variables or “essential” variables, if any, are identified for each failing edit rule and then they are imputed. Frequently, imputing blanks, invalids and essential variables is sufficient for a household to pass the edits. That, along with the techniques developed for the 1996 Census, makes for a very efficient imputation algorithm. The prototype uses highly efficient methods to perform editing but, due to a lack of development time, does not use the 1996 approach of retaining only edit rules that one or more imputation actions could fail.

23. This prototype software was sent to the national statistical agencies of Britain, Brazil and Italy at their request. The Brazilians and Italians have successfully used it to process small quantities of demographic data from their Censuses. The Brazilians are considering the use of NIM for their next Census. The Italians are planning a larger test with the NIM where they will introduce random errors and then assess the quality of the imputation actions.

VIII. EXTENDING NIM TO NUMERIC E&I

24. NIM must be designed for the 2001 Census to be easily generalizable such that it is able to impute all Census variables for the 2006 Census. For this reason, it is useful to consider how to extend it to carry out imputation for a large number of numeric variables. NIM could then process much larger numeric E&I problems than can be handled by systems such as GEIS because of the great computational cost in determining the minimum number of variables to impute under the Fellegi/Holt algorithm. The restrictions on the type of edits that NIM could handle would also be much less than those of GEIS.

25. A typical small set of GEIS edits expressed in DLT form is given in Table 9 in validity rule form and in Table 10 in conflict rule form. The j^{th} numeric proposition below ($j = 1$ to 4) takes the form $V_j = \sum_i B_{ji} V_{ai} - c_j \leq 0$ while V_{ai} , $i = 1$ to I , represent the value for the I numeric variables being edited after some have been imputed, and B_{ji} and c_j represent constants. In this section, the numeric edits, in the form of the four conflict rules of Table 10 will be discussed. NIM, however, can be extended to handle DLTs with numeric propositions with any pattern of Y's and N's. The same approach as outlined for the imputation of a mixture of qualitative and numeric variables in Sections 3 and 7 will be applied here. A nearest neighbour will be found and blanks/invalids plus essential variables will be imputed. If the record still fails, edit rules will be dropped that none of the possible imputation actions would fail. Then the minimum number of imputation actions possible will be generated by discarding any that are not NM CIA, any that are not essentially new or any that will continue to fail specific edits regardless of the additional variables imputed.

Table 9: Validity Rules With GEIS Propositions

$V_1 \leq 0$	Y
$V_2 \leq 0$	Y
$V_3 \leq 0$	Y
$V_4 \leq 0$	Y

Table 10: Conflict Rules With GEIS Propositions

$V_1 \leq 0$	N	-	-	-
$V_2 \leq 0$	-	N	-	-
$V_3 \leq 0$	-	-	N	-
$V_4 \leq 0$	-	-	-	N

26. With qualitative variables, a large number of variables can often be immediately discarded from consideration for imputation because they take on identical values for the nearest neighbour and the failed edit record. With numeric variables, it is quite possible that most if not all of them will have at least slightly different values when the nearest neighbour and the failed edit record are compared. This, initially, makes it appear more difficult to identify if a numeric proposition is always true or always false because the number of possible imputation actions is astronomical. Also, the likelihood of any essential to impute variables being present in the proposition might seem remote given the large number of variables. The fact, however, that we are dealing exclusively with numeric variables in these propositions allows them to be evaluated very rapidly. This can be seen by first noting that the imputed value V_{ai} , for the i^{th} variable can be written as

$$V_{ai} = \ddot{a}_i V_{pi} + (1 - \ddot{a}_i) V_{fi} = \ddot{a}_i (V_{pi} - V_{fi}) + V_{fi}$$

where V_{fi} represents the value from the failed edit record, while V_{pi} represents the value from the nearest neighbour, and \ddot{a}_i represents an indicator variable such that $\ddot{a}_i = 1$ if the i^{th} variable is imputed and $\ddot{a}_i = 0$ if the i^{th} variable is not imputed.

The function V_j in the j^{th} proposition can then be written as

$$V_j = \sum_i B_{ji}^* \ddot{a}_i - c_j^*$$

where $B_{ji}^* = B_{ji}(V_{pi} - V_{fi})$ while $c_j^* = c_j - \sum_i B_{ji} V_{fi}$.

It should be noted that V_j as a function of \ddot{a}_i allows the conflict rules to be rapidly evaluated as variables are sequentially imputed.

27. The concept of essential to impute variables can be generalized in the case of conflict rules involving solely numeric variables. Assume that the initial imputation action, after imputing blank and invalid variables, fails the j^{th} conflict rule, i.e. $V_j^0 > 0$ where V_j^0 represents the value of V_j for the initial imputation action. Assume, in addition, that imputation actions which can be generated from the initial imputation action and which also pass the j^{th} conflict rule always have certain variables imputed (not counting those that were imputed because they were blank or invalid). These variables, if any, will be called the essential to impute variables for the j^{th} conflict rule. The essential to impute variables can be easily identified by first calculating $\min V_j = V_j^0 + \sum_i B_{ji}^*$ where $\sum_i B_{ji}^*$ represents the summation of those values of B_{ji}^* which are negative but only for variables not already imputed because they were blank or invalid. It is known that $\min V_j \leq 0$, because the donor record passes the j^{th} edit. Then we will determine for each variable with a negative B_{ji}^* if $\min V_j - B_{ji}^* > 0$. If this relationship holds, that variable is one of the essential to impute variables because removing it will cause any other imputation action based on the other variables to fail the j^{th} conflict rule.

28. Assume that the initial imputation action still fails the edits after imputing blank, invalid and essential variables. We wish to determine if the j^{th} conflict rule can be discarded because $\max V_j \leq 0$ where $\max V_j$ represents the maximum value possible for V_j based on the initial imputation action plus any imputation actions that can be generated from it. Having this relationship hold means that there is no imputation action that will fail the j^{th} edit for that failed record/donor pair. It is easy to see that $\max V_j = V_j^0 + \sum_{i+} B_{ji}^*$ where $\sum_{i+} B_{ji}^*$ represents the summation of those values of B_{ji}^* which are positive but only for variables not already imputed because they were blank, invalid or essential. Later in processing, after a number of imputation actions have been generated, it can be determined if $\min V_j > 0$ for all imputation actions that can be generated from a specific imputation action. In that situation, that specific imputation action can be discarded because all imputation actions that can be generated from it will fail the j^{th} conflict rule.

29. These generalizations still have to be programmed and tested with a numeric E&I problem to determine how efficient the algorithm would be. Conceptually, however, they appear very promising.

IX. CONCLUDING REMARKS

30. Having been successfully implemented for the 1996 Census, the NIM software is being generalized to process a wider range of variables for the 2001 and 2006 Canadian Censuses. A prototype version of NIM has been created which has allowed us to experiment with certain enhancements plus it has allowed statistical agencies in two other countries to experiment with NIM. NIM could be extended to allow minimum change donor imputation to be done efficiently for imputation problems involving a large number of numeric variables though this must be confirmed numerically. This generalization should allow NIM to be utilized by a wide range of surveys.

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.

Bankier, M., Luc, M., Nadeau, C. and Newcombe, P. (1995), "Additional Details on Imputing Numeric and Qualitative Variables Simultaneously", Proceedings of the Section on Survey Research Methods, American Statistical Association, 287-292.

Bankier, M., Luc, M., Nadeau, C. and Newcombe, P. (1996), "Imputing Numeric and Qualitative Census Variables Simultaneously", Proceedings of the Survey Research Methods Section, American Statistical Association, 90-99.

Bankier, M., Houle, A.M., Luc, M. and Newcombe, P. (1997), "1996 Canadian Census Demographic Variables Imputation", Proceedings of the Survey Research Methods Section, American Statistical Association, 389-394.

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.