A Comparison Study of ACS If-Then-Else, NIM, and DISCRETE Edit and Imputation Systems Using ACS Data*

Bor-Chung Chen, Yves Thibaudeau, and **William E. Winkler**
Bor-Chung Chen, Bureau of the Census, Washington, DC 20233-9100

Key Words: Explicit Edits, Decision Logic Tables, Integer Programming, Optimization

1. Introduction

In any statistical surveys, the information gathered may contain inconsistent, incorrect, or missing data. These erroneous data need to be revised or filled in prior to data tabulations and retrieval. The revisions of the erroneous data should not affect the statistical inferences of the data. The missing data, as well as some inconsistent or incorrect data, are easy to identify while others are not. For those not easily identified, a set of edit rules is needed to specify whether a set of data record is erroneous. One of the important steps of this systematic revision process of the erroneous data is computer editing. The edit rules are traditionally implemented with computer coding of if-then-else structures and many statistical agencies have chosen to adopt these methods. The disadvantages of the if-then-else structures are that they may not be straightforward to develop and may be difficult to write the computer code to implement them. In addition, if there are slight changes in the edit rules or survey form, the software may not be reusable, which will cause thousands of lines of code to be rewritten and debugged.

In this paper, we will compare the if-then-else (ITE hereafter) rules with alternative approaches that have potential to improve the data quality of survey data. The alternative approaches are the Fellegi-Holt model based DISCRETE edit system of the U.S. Bureau of the Census and NIM of Statistics Canada. We use the 1999 American Community Survey (ACS) data of 26 states for the comparisons. The ITE rules used are described in the 1999 ACS Edit and Allocation Specifications for Basic Population Variables, which include sex, age, household relationship, marital status, race, and Hispanic origin. Only the the first four variables are included in this study.

The DISCRETE edit system (Winkler and Petkunas [1996]) is designed for general edits of discrete data. It utilizes the Fellegi-Holt model of editing and contains two major components: edit generation and error localization. Fellegi and Holt [1976] provided an underlying basis of developing another implementation of computer edit system. Their methods have the virtues that the logical consistency of the entire set of edit rules can be checked before the survey data become available and that, in one pass through the data, an edit-failed and imputed record can be assured to satisfy all edits. The implementations of the system have had additional advantages over traditional if-then-else rule edit systems because edits reside in easily modified tables and computer code needs no modification. Fellegi and Holt (FH hereafter) described three criteria for imputation:

1. The data in each record should be made to satisfy all edits by changing the fewest possible items of data (*variables* or *fields*).
2. Imputation rules should be derived automatically from edit rules.
3. When imputation is necessary, it is desirable to maintain the marginal and joint frequency distributions of variables.

These three criteria are very important to maintain a high data quality of the survey data when some of them are inconsistent or missing. The first of them is the core portion of the DISCRETE edit system. It is referred to as the *error localization* (EL) problem. In addition to (explicit) edits that are originally defined, FH showed that precise knowledge of implicit edits was needed. *Implicit edits* are those that can be logically derived from explicit edits. FH (Theorem 1) proved that implicit edits are needed for solving the EL problem. FH provided an inductive, existence-type proof to their Theorem 1. Their solution, however, did not deal with many of the practical computational aspects of the problem that, in the case of discrete data, were considered by Garfinkel, Kunnathur, and Liepins [1986] , which improvements were implemented in the current DISCRETE edit system used in this study.

Bankier [1997, see also 2000] introduced a successful method for using (hot-deck) donor imputation that has been used for the 1996 and 2001 Canadian Censuses and will be used for the 2006 Canadian Census. As with other donor imputation systems, the method is dependent on having a large population of high quality donors. Before describing NIM in Section 3, we describe how a corresponding FH edit system that uses hot-deck imputation would work. The FH edit system would determine the minimum number of fields to change. A priori matching rules would be developed to select hot-deck donors from the set of records that satisfy all edits. If there are suitable donors, then imputed fields from the hot-deck donors will maintain the univariate distributions of the respondents. Two difficulties are associated with systems that use hot-deck imputation. The first is that the matching rules may not be as good as they can be. This has been noted as a problem in the 1990 U.S. Decennial Census, the

*This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

1991 Canadian Census, and the 1991 British Census. The second is that there may not be enough suitable donors. The second problem is often not as serious in a census as it is in a smaller survey.

This paper compares results from the second version of the DISCRETE edit system, an earlier version of the NIM system, and the first version of the ACS ITE system. The earlier version of NIM differs somewhat from the current version. The ACS ITE system also has been updated. Our main comparisons are primarily among the imputed values from the ACS ITE system and NIM.

2. Existing ITE Rules Used by ACS

As mentioned in Section 1, the existing if-then-else rules used by ACS are described in *the 1999 ACS Edit and Allocation Specifications for Basic Population Variables (Sex, Age, Household Relationship, Marital Status, Race, and Hispanic Origin)*. The specifications provide the edit for each population variable, including data definitions and edit rules. In this paper, we only study the variables of sex, age, household relationship, and marital status.

The specifications are divided into sections by variables. Each of the variables, sex and age, has its own section. The variables of household relationship and marital status are in the same section. The division into sections by variables has its meaning of making changes on the variables. If an "if" condition is satisfied in the "age" section, the imputation of age, rather than sex or any other variables, will be performed. The nature of the ITE rules combined with the division into section by variables might have different imputation results if the orders of processing the sections are different. Also, there is no guarantee for each edit-failing household passing all edits if only one iteration of the system is performed.

In the edit and imputation process, the data file is initially sorted by state, county, tract, block group, and sequence. When a donor is needed for an edit-failing record, the system searches forward and backward from the record. The search starts within the block group, then within the tract, and the county and state until an appropriate donor is found. If none is found, a value from the matrix associated with that variable is used as the imputed value.

3. Bankier's NIM Methodology

Bankier's NIM proceeds primarily by using donors. Each edit-failing record is matched with a large subset (say 1,000) of records that satisfy all of the edits. The ones, say 20, that have the smallest deviations in terms of the number of fields differing from the edit failing record are retained as the potential donors and are called *nearest neighbors*. To obtain the smallest deviations, NIM first searches, in the imputation group, for those edit passing records $\boldsymbol{a}_p$ that are closest to the edit failing record $\boldsymbol{a}_f$ in terms of the distance,

$$D_{fp} = D(\boldsymbol{a}_f, \boldsymbol{a}_p) = \sum_i \omega_i D_i(a_{fi}, a_{pi}) \qquad (1)$$

where the *weights* $\omega_i \geq 0$ can be given smaller values for variables where it is considered less important that they match, i.e., variables considered more likely to be in error. In this study, all ω_i were set to one. The distance $D_i(a_{fi}, a_{pi})$ between the edit failing record and the edit passing record for the ith field is, for discrete fields,

$$D_i(a_{fi}, a_{pi}) = \begin{cases} 0 & \text{if } a_{fi} = a_{pi} \\ 1 & \text{otherwise} \end{cases}, \quad \text{or,}$$

for continuous fields,

$$0 \leq D_i(a_{fi}, a_{pi}) \leq 1 \qquad (2)$$

in which $D_i(a_{fi}, a_{pi}) = 0$ if $a_{fi} = a_{pi}$ and $D_i(a_{fi}, a_{pi})$ is an increasing function of $|a_{fi} - a_{pi}|$. The form of the distance measure can be different for each type of continuous field as long as it respects the restrictions of (2).

The distance measure, $D_i(a_{fi}, a_{pi})$, for the age variables used in this study is defined as follows.

$D_i(a_{fi}, a_{pi}) =$

$$\begin{cases} 1 & \text{if } |a_{fi} - a_{pi}| \geq m(a_{fi}) \\ 1 & \text{if } a_{fi} \text{ is missing or invalid} \\ 1 & \text{if } a_{fi} \geq 15 \text{ and } a_{pi} < 15 \\ 1 & \text{if } a_{fi} < 15 \text{ and } a_{pi} \geq 15 \\ 1 - \left(1 - \frac{|a_{fi} - a_{pi}|}{m(a_{fi})}\right)^r & \text{otherwise} \end{cases}$$

where $r \geq 0$ is a constant and was set to 0.25 and

$$m(a_{fi}) = \begin{cases} k_1 + \frac{k2(a_{fi} - k_3)}{10} & \text{if } a_{fi} > k_3 \\ k_1 & \text{if } a_{fi} \leq k_3 \end{cases}$$

The parameters k_1, k_2, and k_3 were set to 6, 2, and 30, respectively, in this study. If $D_i(a_{fi}, a_{pi}) = 1$, the two age variables, a_{fi} and a_{pi}, are considered as nonmatching.

Feasible *imputation actions* $\boldsymbol{a}_a$ are then generated from each of the potential donors. Feasible imputation actions are changes to some fields of the edit failing record so that the new imputed record may pass all edits. Then, the feasible imputation actions $\boldsymbol{a}_a$ for each edit failing/passing record pair are identified such that $\boldsymbol{a}_a$ passes the edits and the distance

$$D_{fpa} = \alpha D_{fa} + (1 - \alpha) D_{ap} \qquad (3)$$

is minimized or nearly minimized, where

$$D_{fa} = \sum_i \omega_i D_i(a_{fi}, a_{ai})$$

is the distance between the imputation action and the edit failing record,

$$D_{ap} = \sum_i \omega_i D_i(a_{ai}, a_{pi})$$

is the distance between the imputation action and the nearest neighbor used, and α is a parameter that falls in the range (0.5, 1]. Values of α close to 1 indicate that more emphasis is placed on imputing the minimum number of variables than having the imputed household resemble the donor. The value of α was set to 0.9 in this study. D_{fa} is a measure of how many variables are imputed. D_{ap} is a measure of plausibility.

Any feasible imputation actions with $D_{fpa} = \min\{D_{fpa}\}$ are called *minimum change imputation actions*. Those feasible imputation actions with a D_{fpa} that satisfy

$$D_{fpa} \leq \gamma \min\{D_{fpa}\} \qquad (4)$$

are retained and are called *near minimum change imputation action* (NMCIA), where γ was set to 1.025 in this study. The n, say 5, feasible imputation actions with smallest D_{fpa}, the weighted average of D_{fa} and D_{ap}, are retained. Then one of

these n imputation actions is randomly selected to be the actual imputation action used for the edit failing record.

4. DISCRETE Editing System

We will use the following notations in the brief description of the DISCRETE edit system: $\boldsymbol{a}=(a_1, a_2, \ldots, a_n)$ has n fields. For each i, $a_i \in A_i$, $1 \le i \le n$, where A_i is the set of possible values or code values which may be recorded in Field i. $|A_i| = n_i$. If $a_i \in A_i^o \subset A_i$, we also say

$$\boldsymbol{a} \in \boldsymbol{A}_i^o = A_1 \times A_2 \times \ldots \times A_{i-1} \times A_i^o \times A_{i+1} \times \ldots \times A_n.$$

The code space is $A_1 \times A_2 \times \ldots \times A_n = \boldsymbol{A}$.

The objective of error localization is to find the minimum number of fields to change if a record fails some of the edits. It can be formulated as a set covering problem. Let $\bar{E} = \{E^1, E^2, \cdots, E^m\}$ be a set of edits failed by a record $\boldsymbol{y}$ with n fields, consider the set covering problem:

$$\begin{aligned} &\text{Minimize} && \textstyle\sum_{j=1}^{n} c_j x_j \\ &\text{subject to} && \textstyle\sum_{j=1}^{n} a_{ij} x_j \ge 1, \quad i = 1, 2, \cdots, m \quad (5) \end{aligned}$$

$$x_j = \begin{cases} 1, & \text{if field } j \text{ is to be changed;} \\ 0, & \text{otherwise,} \end{cases}$$

where

$$a_{ij} = \begin{cases} 1, & \text{if field } j \text{ enters } E^i; \\ 0, & \text{otherwise,} \end{cases}$$

and c_j is a measure of *confidence* in field j. A small value of c_j indicates that the corresponding field j is considered more likely to be in error. In this study, c_j was set to 3.50 for the sex variable, 5.20 for the household relationship variable, 2.10 for the marital status variable, and 2.07 for any of the age comparison variables. We need to get $\bar{E}$ from a *complete* set of edits to obtain a meaningful solution to (5). A complete set of edits is the set of explicit (initially specified) edits and all essentially new implied edits derived from them.

If $\boldsymbol{x}$ is a prime cover solution to (5) and $K = \{r \mid x_r = 1\} \subset \{1, 2, \cdots, n\}$, then for each $k \in K$ we may change the value of field f_k to a value from

$$B_k^* = \overline{\bigcup_{j \in J} A_k^j} = \bigcap_{j \in J} \overline{A_k^j},$$

where

$J = \{j \mid 1 \le j \le m,\ f_k \text{ is an entering field of } E^j\}$. The new imputed record $\boldsymbol{y}_1$, which has different value of $f_k\ \forall\ k \in K$ from the record $\boldsymbol{y}$, will pass all edits. Note that $B_k^* \neq \emptyset$. If B_k^* were a empty set, then $\bigcup_{j \in J} A_k^j$ would be equal to A_k and an essentially new implicit edit would have been generated and included in the set of $\bar{E}$.

To obtain a *complete* set of edits, implicit edits are needed. Implicit edits may be implied logically from the initially specified edits (or explicit edits). Implicit edits give information about explicit edits that do not originally fail but may fail when a field in a record with an originally failing explicit edit is changed. For a detailed description of edit generation, see Fellegi and Holt [1976].

Therefore, the DISCRETE edit system consists of two components: the edit generation program and the error localization program. To apply the system to the 1999 ACS data set, we need two additional components: the age comparison program and the pre-edit program. The pre-edit program is described in Section 6. The age comparison program is based on new age comparison variables described in Chen and Winkler [2002], which has a better performance than the one described in Chen, Winkler, and Hemmig [2000] .

5. Pre-Edits

Some missing fields in a record can be logically derived from other non-missing fields. These type of edits, also referred to as logical edits, is called pre-edits. Other pre-edits common to the three systems compared in this study are (1) identify the householder and spouse if present; (2) perform household relationship pre-edits; (3) perform age and date of birth pre-edits and the consistency checks between age and date of birth; and (4) perform marital status pre-edits.

The first person in a household is usually identified as the householder. It is also possible that a parent becomes the householder, in which the household relationship of the other persons in the same household has to be changed, in which the parent who becomes the householder is considered the *first Father/Mother*. The spouse or spouse-equivalent, such as unmarried partner, roommate, or housemate, is also identified if there is one. If there is more than one spouse or spouse-equivalent, the sequence of spouse, unmarried partner, roommate, and housemate is used to be the second person. The duplicates will be changed to *other nonrelative*.

One of the important characteristics of NIM is that it requires a very high percentage of qualified donors. The set of imputed values of an edit failing record has to be from a single donor. Therefore, the importance of pre-edits in NIM is illustrated in Table 1, which lists the percentage of failed records for different household sizes with and without pre-edits. Table 1 also indicates that there is a very high percentage of edit failing households without pre-edits when the household size becomes large and it drops significantly with pre-edits. A high percentage of edit failing households can mean that there is not enough donors to preserve the statistical properties of the survey data set.

Table 1: Percentage Failed with NIM.

HHD Size	*Total HHDs*	*w/o pre-edits % Failed*	*w/ pre-edits % Failed*
3	16954	40.03	27.47
4	14258	45.22	33.56
5	6742	50.86	39.20
6	2129	92.44	55.85
7	719	95.27	60.92
8	319	96.55	64.89
9	150	96.67	64.00
Total	41271	47.90	33.96

The DISCRETE edit generation and the age comparisons are two major steps before the actual production of the data is performed. The pre-edit step for DISCRETE is also the preparation for fitting the Fellegi-Holt model and performing the production when the data are available. One other purpose of the pre-edits is to convert each of the households into a three-person household. Here, we assume that

there are at most three generations living in a household so that each household is converted into a three-person household, in which the householder and the spouse (or spouse-equivalent) if present are, respectively, the first and second members. The third member will be one of the others.

6. The Specified Variables and Their Values

In this report, we interchangeably use *variable* and *field* in a record. The variables used in the ITE rules are part of the computer program. Their definitions are detailed in *the 1999 ACS Edit and Allocation Specifications for Basic Population Variables*. In NIM system, we specified three coded variables for each person in the ACS households: `SEXU`, `RELANU`, and `MARSTU` representing sex, household relationship, and marital status, respectively. The fourth variable is `AGEU`, which has a value between 0 and 115. Therefore, the number of variables in a record depends on the household size, in which each household is a record.

To identify the explicit edits of DISCRETE system for this study, we assume that each household has at most 3 members, in which the first member is the householder and the second member is the spouse of the householder if there is one.Therefore, the first nine of the 15 fields identified are sex, household relationship, and marital status for the three members in the household: `SEXU11` (meaning the first person's sex), `RELANU11`, `MARSTU11`, `SEXU22`, `RELANU22`, `MARSTU22`, `SEXU33`, `RELANU33`, and `MARSTU33`. The other 6 fields are for the age comparison condition variables.

In the age comparison, each time when a new age restriction appears in one of the if-then-else rules in the 1999 ACS Edit and Allocation Specifications, a temporary age comparison condition variable is defined. A temporary age comparison condition variable is an inequality of the form:

$$a_1x_1 + a_2x_2 + a_3x_3 > b, \qquad (6)$$

where a_i $(i = 1, 2, 3)$ is one of the three values: -1, 0, and 1, and x_i is the ith person's age. There are three possible values for each of the age comparison condition variables: 1 if (6) is true; 2 if false; and 3 if unknown. We identified 41 temporary age comparison condition variables of Inequality (6) for this study.

The 41 temporary Age comparison condition variables can be converted into six variables with the form (see Chen and Winkler [2002]):

$$a_1x_1 + a_2x_2 + a_3x_3, \qquad (7)$$

where $(a_1,\ a_2,\ a_3)$ is one of the following triples: (0,0,1), (0,1,0), (0,1,-1), (1,0,0), (1,0,-1), and (1,-1,0). The six variables are then fit to the Fellegi-Holt model described in Section 4. The formulation significantly reduced the size of the set covering problem of the edit generation and the error localization.

7. Edit Rules

The edit rules are specifications that describe what types of data combinations for the fields of a record are allowed or not allowed. Therefore, there are two types of edit rules: validity rules and conflict rules. The validity rules specify certain types of data combinations are allowed and the conflict rules specify those that are not allowed. All of the three systems in this study specify the edit conflict rules. One example of the edit rules for the ITE system is given in Table 2. The edit rule in this example is the "Universe" and "If" portions of the specification. They have to be converted into a computer code, which is part of the executable. When the edit rules are changed, the program has to be rewritten. It makes programming from scratch absolutely necessary if a new survey and new edit rules are specified.

Table 2: If-Then-Else Edit Specification.

Universe	Person 2+ and *Relationship* is Husband/wife;
If...	*Marital status* is Widowed, divorced, separated, or never married;
Then...	Make *Marital status* = Married; tally TP4(4); set allocation flag.

The NIM system uses decision logic tables (DLT) to store the edit rules. Unlike the ITE system, the DLTs are input to the NIM program. The changes of the edit rules only requires the changes of the DLTs and the NIM program itself is not changed. A DLT is a matrix where the first column is a list of propositions (such as `RELANU(03) = MOTHER`) followed by columns of Y's, N's and spaces that each represent an edit rule. An example of a DLT is given in Table 3. The first column of the Y's, N's, and spaces represents the edit rule described in Table 2. A total of 16 DLTs has been identified for this study. The 16 DLTs consist of 210 propositions and 121 edit rules. Each of the propositions and edit rules directly came from *the 1999 ACS Edit and Allocation Specifications*.

Table 3: DLT of Edit Rules with NIM.

```
RELANU(01) = PERSON1          ;Y;Y;Y;Y;Y;Y;
RELANU(02) = HUSBAND_WIFE     ;Y;Y;Y;Y;Y; ;
SEXU(01) = SASMIS             ; ;Y;Y; ; ; ;
SEXU(02) = SASMIS             ; ; ; ;Y;Y; ;
SEXU(01) = MALE               ; ; ; ;Y; ; ;
SEXU(01) = FEMALE             ; ; ; ; ;Y; ;
SEXU(02) = MALE               ; ;Y; ; ; ; ;
SEXU(02) = FEMALE             ; ; ;Y; ; ; ;
MARSTU(02) = NOW_MARRIED      ;N; ; ; ; ; ;
AGEU(01) > -1                 ; ; ; ; ; ;Y;
AGEU(01) < 15                 ; ; ; ; ; ;Y;
```

The DISCRETE edit system uses edit tables. An edit table is a set of edit rules that are listed with an easily understandable expression. The edit rule in Table 2 is translated into the normal form of the edit:

$$A_1 \times \{1\} \times A_3 \times A_4 \times \{2\} \times \{2,3,4,5\} \times A_7 \times \cdots \times A_{15} = F$$

with $A_2^o = \{1\}$ (`RELANU11`), $A_5^o = \{2\}$ (`RELANU22`), and $A_6^o = \{2,3,4,5\}$ (`MARSTU22`). Fields 2, 5, and 6 are called *entering fields* of the edit because $A_2^o \neq A_2$, $A_5^o \neq A_5$, and $A_6^o \neq A_6$. The edit places restrictions on the values that fields 2, 5, and 6 can assume. The other fields are called *uninvolved* of the edit. Therefore, it is sufficient to identify an edit with its entering fields and their values as it is with the input format of the DISCRETE program:

```
Explicit edit # 25: 3 entering field(s)
  RELANU11      1 response(s):  1
  RELANU22      1 response(s):  2
  MARSTU22      4 response(s):  2  3  4  5
```

Like the NIM system, the DISCRETE system has

the edit table as input to the program. Any changes to the edit rules require the edit table changes only, there is no need to change the DISCRETE program code.

A total of 141 explicit edits has been identified for this study. Seventy-four of them directly came from *the 1999 ACS Edit and Allocation Specifications*. The age comparison program identified the other 67 explicit edits, each of which is a contraction condition within a subset of the 6 age comparison variables discussed in Section 6.

8. Statistical Comparisons

One of the important criteria raised by Fellegi and Holt [1976] was to maintain the frequency distributions of variables when imputation is necessary as described in Section 1. In this section, we compare the frequency distributions of the imputed data among the three systems to that of the edit-passing households. We intend to identify the system that has a "closer" frequency distribution to that of the edit-passing households. The edit-passing households are the "clean" survey data that would represent the survey sample which, in turn, is used to draw the statistical inferences for the population. Therefore, we will use the edit-passing households as a benchmark to determine which system has a "better" imputation results. We will have 4 univariate frequency distributions: sex, age, household relationship (hhr), and marital status (ms); and 6 bivariate frequency distributions: sex-age, sex-hhr, sex-ms, age-hhr, age-ms, and hhr-ms.

We define the "closeness" measurement between the sets of the imputed households and the edit-passing households as the sum of squared deviations between their frequency distributions:

$$\sum_{i=1}^{n} (x_i - y_i)^2, \qquad (8)$$

where n is the number of categories or the number of all possible valid values of a variable; x_i and y_i are the proportions of individuals in the edit-passing and imputed households, respectively, who belong to category i. The valid age is bewteen 0 and 115 that is divided into 23 categories with 5 years in each category except the last one which has 6 years.

In the comparisons among the three systems, a small value of the sum of squared deviations of (8) of an imputed data set would represent a "look alike" frequency distribution of the edit-passing households. Therefore, we would like to have an imputation system that provide a smaller value of (8). Table 4 lists the values of (8) for the ITE and NIM systems by variables and household sizes. The row of "sum" is the sum of the values from rows "3" to "9" representing the aggregate measurement of each of the univariate and bivariate frequency distributions. From Table 4, it is clear that NIM outperforms the existing ITE system. The data of the DMB (DISCRETE Model-Based) are available in Chen, Thibaudeau, and Winkler [2002].

9. Comparisons of Imputed Results

In this section, we discuss the comparisons of the imputed results of the edit failing households from the ITE and NIM systems. According to Table 5, the total number of households imputed for this study is 13,844. There are 10,689 households, or 77.2%, that have exactly the same imputed results with the ITE rules and NIM. The other 3,155 households, or 22.8%, have at least one imputed values disagreed.

Table 4: Comparisons of Sum of Squared Deviations.

vars	sex		ms	
size	ITE	NIM	ITE	NIM
3	0.0012	0.0007	0.0021	0.0009
4	0.0014	0.0011	0.0019	0.0001
5	0.0002	0.0000	0.0019	0.0000
6	0.0014	0.0018	0.0128	0.0012
7	0.0001	0.0002	0.0175	0.0005
8	0.0014	0.0003	0.0171	0.0010
9	0.0113	0.0095	0.0188	0.0046
sum	0.0170	0.0136	0.0721	0.0083
vars	age		hhr	
size	ITE	NIM	ITE	NIM
3	0.0027	0.0011	0.0169	0.0021
4	0.0030	0.0015	0.0216	0.0025
5	0.0043	0.0017	0.0193	0.0029
6	0.0054	0.0005	0.0123	0.0006
7	0.0100	0.0009	0.0102	0.0014
8	0.0041	0.0018	0.0027	0.0005
9	0.0175	0.0046	0.0120	0.0038
sum	0.0470	0.0121	0.0950	0.0138
vars	sex-ms		sex-age	
size	ITE	NIM	ITE	NIM
3	0.0016	0.0008	0.0017	0.0007
4	0.0015	0.0006	0.0017	0.0009
5	0.0010	0.0001	0.0024	0.0009
6	0.0068	0.0014	0.0030	0.0006
7	0.0088	0.0004	0.0055	0.0010
8	0.0090	0.0006	0.0035	0.0021
9	0.0183	0.0110	0.0133	0.0066
sum	0.0470	0.0149	0.0311	0.0128
vars	sex-hhr		ms-age	
size	ITE	NIM	ITE	NIM
3	0.0096	0.0023	0.0032	0.0021
4	0.0113	0.0019	0.0043	0.0019
5	0.0101	0.0015	0.0048	0.0019
6	0.0068	0.0013	0.0049	0.0005
7	0.0060	0.0024	0.0091	0.0008
8	0.0024	0.0009	0.0042	0.0018
9	0.0125	0.0099	0.0148	0.0048
sum	0.0587	0.0202	0.0453	0.0138
vars	ms-hhr		age-hhr	
size	ITE	NIM	ITE	NIM
3	0.0179	0.0034	0.0051	0.0015
4	0.0226	0.0027	0.0070	0.0018
5	0.0212	0.0034	0.0079	0.0024
6	0.0118	0.0011	0.0043	0.0007
7	0.0098	0.0022	0.0062	0.0016
8	0.0050	0.0019	0.0024	0.0021
9	0.0171	0.0041	0.0123	0.0047
sum	0.1054	0.0188	0.0452	0.0148

In the following, we list some disagreements that several imputed households are still problematic after ITE and/or NIM imputations. See Chen, Thibaudeau, and Winkler [2002] for other disagreements.

1. When a married *unmarried partner*, *child*, *parent* has similar age of the married householder and the spouse is missing in the household, the ITE system calls this person *roomer/boarder*, *brother/sister*, *other relative*, or *other relative* and NIM calls him/her *spouse*.
2. There was an example that shows the ineffective sequential edit system, such as the ITE system. After imputing a value for the *marital status* of the second person, the spouse, of a household,

the ITE system made an unnecessary change of the third person's (a daughter) age of 12 to 24, that fails the edit of the householder's age of 36 must be at least 15 years older than a child.

3. We have an example that the *minimum number of fields to change* may not be a *reasonable* imputation for NIM, in which the third person's relationship is imputed with the value of *son* (one-field inputation; the unedited value was missing). The ITE rules also change the second person's relationship to *spouse* (two-field imputation; the unedited value was *other nonrelative*).
4. We have another example that the *nearest neighbor imputation* may not be a *reasonable* imputation for NIM, in which the donor provides the third person's relationship of *other relative* instead of *daughter* like the ITE rules provide (the unedited value was missing; the householder's age was 41 and the third's age was 9).

Table 5: Agreed and Disagreed Imputations: ITE vs. NIM.

HHD Size	*Total # of HHDs Imputed*	*Number of Agreed*	*Number of Disagreed*
3	4658	3564	1094
4	4774	3958	816
5	2643	2007	636
6	1028	720	308
7	438	269	169
8	207	117	90
9	96	54	42
Total	13844	10689	3155

10. Discussion and Summary

The results of this study indicate that NIM and DISCRETE always identify the same edit-passing and edit-failing household records. There are many cases that the If-Then-Else rules could not make a edit-failing household record to pass all edits. One of the important criteria raised by Fellegi and Holt was to maintain the frequency distributions of variables when imputation is necessary. Therefore, we also compared the frequency distributions of the imputed data among the systems to that of the edit-passing households. We intended to identify the system that has a "closer" frequency distributions of the imputed households to that of the edit-passing households. The edit-passing households are the "clean" survey data that would represent the survey sample which, in turn, is used to draw the statistical inferences for the population. Therefore, we used the edit-passing households as a benchmark to determine which system has a "better" imputation results. We defined the "closeness" measurement between the sets of the imputed households and the edit-passing households as the sum of squared deviations between their frequency distributions. The initial results indicate that NIM outperforms the existing If-Then-Else system. Another advantage of NIM and DISCRETE over the If-Then-Else rules is that the computer code does not need to be rewritten from a survey to another when the edit rules change.

With the larger household sizes, it is often difficult to have a sufficient number of suitable donors for the hot-deck imputation used by ACS ITE and NIM. Model-based imputation methods can also have difficulty when there are many more variables in a record. Our application shows that the ACS ITE and NIM produce different imputations. No system can produce perfect imputations. A few of the imputations produced by NIM seem more plausible than the imputations from the earlier version of ACS ITE. The longer version of this paper will deal more fully with imputation differences.

References

Bankier, M. (1997) Documentation of the New NIM Prototype. Social Survey Methods Division Report, Statistics Canada, Ottawa, Dated September 7, 1997.

Bankier, M. (2000) Imputing Numeric and Qualitative Variables Simultaneously. Research Report, Statistics Canada, Ottawa.

Chen, B., Thibaudeau, Y., and Winkler, W. E. (2002) A Comparison Study of ACS If-Then-Else, NIM, and DISCRETE Edit and Imputation Systems Using ACS Data. Research Report *forthcoming*, Statistical Research Division, Bureau of the Census, Washington, D.C.

Chen, B. and Winkler, W. E. (2002) An Efficient Formulation of Age Comparisons in the DISCRETE Edit System. Research Report Computing 2002-02 Statistical Research Division, Bureau of the Census, Washington, D.C.

Chen, B., Winkler, W. E., and Hemmig, R. J. (2000) Using the DISCRETE Edit System for ACS Surveys. Research Report RR2000/03, Statistical Research Division, Bureau of the Census, Washington, D.C.

Fellegi, I. P. and Holt D. (1976) A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71:17–35.

Garfinkel, R. S., Kunnathur, A. S., and Liepins, G. E. (1986) Optimal Imputation of Erroneous Data: Categorical Data, General Edits. *Operations Research* 34:744–751.

Winkler, W. E. and Petkunas, T. F. (1996) The DISCRETE Edit System. Statistical Research Division Research Report, Bureau of the Census.