WP No. 7

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

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

Invited Paper

Submitted by [Bureau of the Census, USA][1]

ABSTRACT

This paper does a comparison of various edit and imputation systems. The ACS edit and imputation system uses classical if-then-else rules. The Nearest Neighbour Imputation Method of Statistics Canada (Bankier 2000) provides an edit-table-based method that automatically finds nearly optimal hot-deck matching rules and imputes data in an efficient manner so that the resultant records typically satisfy edits. The DISCRETE edit and imputation system is a Fellegi-Holt system that is table-based that determines the minimal number of fields to change (Winkler 1997, Chen 1998) and provides model-based imputation (e.g., Thibaudeau 2002).

[1] Prepared by [bor.chung.hilary.chen@census.gov , yves.thibaudeau@census.gov, william.e.winkler@census.gov].

I. Introduction

1. In any statistical survey, the data items may be inconsistent, incorrect, or missing. To facilitate valid statistical inference using standard methods and software, the values of the fields (variables) associated with the erroneous data need to be revised or filled in. Individuals with subject matter expertise often define edit rules that the data must satisfy. The edit rules assure that the amount of inconsistent and incorrect data are minimized. In modern processing environments, edit/imputation rules are often implemented in software to assure greater consistency and to minimize manual review and corrections.

2. There are several methods for edit/imputation. The classical method implemented by most statistical agencies is to write software that implements if-then-else (ITE) rules for edit and imputation. The classical method has the disadvantage that it is often difficult to implement hundreds or thousands of if-then-else rules in maintainable code. If a survey form is changed or edit restraints are modified, then it may take considerable effort to modify the code. In many instances, statistical agencies have concluded that it is more efficient to rewrite the edit/imputation software in its entirety. More efficient edit methods are based on models in which edit rules are contained in tables. The tables are easily modified. Source code does not need modification. There are two variants of the table-based methods. The first is the Nearest-Neighbor Imputation Method (NIM) that is primarily based on hot-deck imputation. The second is based variants of the model of statistical data editing due to Fellegi and Holt (1976). Our specific variant uses methods of editing discrete data due to Winkler (1995, 1997), Chen (1998), and Chen and Winkler (2002) and imputation methods due to Thibaudeau (2002). The methods are referred to as the DISCRETE Model-Based (DMB) methods.

3. In this paper, we compare the three methods using data from the American Community Survey (ACS) at the U.S. Bureau of the Census. The ACS is a large continuous measurement survey of individuals in households for the entire U.S. To facilitate our comparisons, we only consider the fields associated with sex, age, household relationship, and marital status. We use 1999 ACS data from 26 states.

4. The outline of this paper is as follows. In section two, we provide background on editing and an overview of editing models. In section three, we give an overview of the ACS data and the various forms of edits that are needed for the systems that we compare. In section four, we describe a pre-edit program that puts the data is form that is most suitable for used in the ITE, NIM and DMB systems. In section five, we cover a sophisticated method of that summarizes age information within a household and applies it in editing. The age-summarization method generalizes Chen, Hemmig, and Winkler (2001). In sections six, seven, and eight we provide overviews of the ACS if-then-else (ITE) system, the NIM system (Bankier, 1997, 2000), the DMB system that includes edit methods based on the Fellegi-Holt model in the DISCRETE edit system (Winkler 1997, Chen 1998, Chen and Winkler 2002) and model-based imputation due to Thibaudeau (2002). In section nine, we compare frequency distributions of the subset of ACS data that passes edits with the entire sets of data that have been passed through the ITE, NIM, and DMB systems, respectively. In section ten, we provide additional details of the imputations. We also provide examples of the small number of anomalies observed with each of the three systems. The eleventh section is discussion and the final section is a summary.

II. Background on Editing and Editing Models

5. Edit-imputation methods are designed to create data in which missing data values of fields (variables) are filled in and which contradictory values of in individual fields are changed to those that are no longer contradictory. As an example of an edit of a single field, we might specify that ages must be between 0 and 115 years. In comparisons of two fields, we might specify that a married person must be 15 or more years of age or that a parent must be 15 years older than a child. Edits of single fields are best dealt with during a preprocessing stage called pre-editing. Lookup tables and other straightforward methods are typically used. An edit places restrictions on the values that can be placed in certain fields or combinations of fields. A record fails an edit if the values of the record correspond to the proscribed values of the fields in the edit. For a record to not fail an edit, one value in at least one field associated with the edit must be change.

6. Edits of multiple fields are more difficult to deal with for several reasons. First, in traditional if-then-else systems, the set of explicitly defined edits might be logically inconsistent in the sense that no record could satisfy all of the edits. This is particularly a problem with large survey situations having hundreds of rules and thousands of lines of code. Second, it might be quite difficult to write and maintain several thousand lines of code associated with a large edit situation. Third, there was often no guarantee that a record passing through if-then-else code would satisfy all edits. The main difficulty can be that, after imputing a set of fields, a record might fail an edit that it did not originally fail. Because the editing and imputing of the original record changed at least one field in each failing edit, none of the edits in the set of original failing edits should fail.

7. Fellegi and Holt (1976, hereafter FH) introduced a mathematical model that was intended to solve the three difficulties with the traditional if-then-else methods. The algorithms use the mathematical restraints of the edit rules to determine the logical consistency of the edit rules prior to the receipt of data. The edits reside in easily maintained tables. Source code (particularly mathematical algorithms) does not need to be rewritten as the survey form or edit rules are modified. The theory results in a global optimization that assures that records are "corrected" in one pass. Sequential-hierarchy methods associated with if-then-else rules are unable to assure "corrections" of records even with multiple passes through a record.

8. The goals of Fellegi and Holt for the methodology were defined in three criteria.

   (a) The data in each record should be made to satisfy all edits by changing the fewest possible items of data (*variables* or *fields*).

   (b) Imputation rules should be derived automatically from edit rules.

   (c) When imputation is necessary, it is desirable to maintain the marginal and joint frequency distributions of variables.

9. The first criterion is referred to as *error localization* (EL). EL involves determining the minimum number of fields to impute so that an edit-failing record will satisfy edits. FH showed that implicit edits are needed for solving the EL problem. *Implicit edits* are those edits that are logically derived from the explicitly defined edits. In work prior to FH, individuals would "correct" (i.e., change values

in) fields of failing explicit edits only to discover that the resultant record would fail explicit edits that the original, unchanged record did not fail. FH (Theorem 1) demonstrated that if implicit edits were available, then a record could be "corrected" in one pass so the new record failed no edits.

10. FH methods have been implemented in three ways. First, implicit edits are computed prior to the receipt of survey data (e.g., Winkler 1995; Winkler 1997; Chen 1998). Availability of the implicit yields very fast EL systems (100-1000 records per second). Availability of all implicit edits assure that all records can be error localized automatically. Automatic correction is an advantage in large survey situations having a minimum of hundreds of thousands of records. In most situations, implicit-edit are available because they can be generated in between 6 minutes and 24 hours. With some large labor force surveys, depending on the speed of the algorithms, edit generation may need 1-800 days. Second, integer-programming methods can be used to solve the EL problem directly without computing implicit edits. These methods apply branch-and-bound, cardinality-constrained Chernikova, or Fourier-Motzkin algorithms. The methods limit the amount of computation by only considering the easiest records that require the fewest fields to change or the least amount of computation up to an upper bound. Records that are not error localized must be manually reviewed and corrected. The Chernikova methods (Kovar and Schopiu-Kratina 1989) and the Fourier-Motzkin methods (DeWaal 2000) are known to find failing implicit edits as part of the error localization. Because of the the extra computation, the methods can be slow (0.1 to 1 second per record). They are most suitable for sample surveys having less than 100,000 records. The third method is the Nearest-Neighbor Imputation Method (NIM) introduced by Bankier (1991, see also 2000). NIM matches an edit-failing record against a large set of edit- passing records to obtain a small number of donor records. The fields that differ between the edit-passing and each donor are evaluated in the sense of which ones should be changed so that the resultant changed records still satisfies edits. In the situations where many suitable donor records are available, NIM can be shown to be consistent with an extended version of the FH model (Winkler and Chen 2002). The first and third methods are often more suitable for censuses having millions of records.

11. In the empirical comparisons of this paper, we will compare the ACS ITE system, NIM, and the DMB system that combines the discrete editing methods (Winkler 1997, Chen 1998, Winkler and Chen 2002) with model-based imputation methods (Thibaudeau 2002). The details of the applications of the methods will be covered in separate sections following the description of the data and the set of edits.

III. ACS Data and Edits

12. In this section, we provide a description of the ACS data and a summary of the types of edits that are applied to it. The ACS data that we use consist of households of persons with the four fields relationship to head of household, sex, marital status, and age. The data are subdivided by household size. We consider eight groups corresponding to household sizes between 2 and 9. The chief difficulties with this data are editing and imputing ages of persons within households. The ACS data are from the sample of 78391 households for 26 States from 1999.

13. Values of the relationship field are given in Tables 1 and 2 in the form needed for NIM and in Table 3

for the form needed by DMB that also corresponds to the form needed for ITE. For all three systems, the sex field takes values male, female, and missing shown in the first column in Table 3. The marital status field takes six values shown in the last column of Table 3 for DMB. The relationship field takes the 16 values shown in Table 3 for NIM. In applying the edits in the DMB, we divide each household of size greater than three into groups containing the householder and two other individuals. The groupings allow us to deal with three generations of ages within a household. The grouping allows us to deal efficiently with the combinatorial explosion of implicit edits that would be needed otherwise. This type of heuristic is not needed for the other systems.

Table 1: Valueset of the three coded variables with NIM.

Variable	Values and Response Classes	Notes
SEXU	MALE	
	FEMALE	
	SASMIS	Unknown or invalid value
RELANU	*PARTNER	A response class
	*RELATIVE	A response class
	PERSON1	Householder
	HUSBAND_WIFE	
	SON_DAUGHTER	
	BROTHER_SISTER	
	FATHER_MOTHER	
	GRAND_CHILD	
	IN_LAW	
	OTHER_REL	Other relative
	ROOMER	Roomer/boarder
	HOUSEMATE	Housemate/roommate
	UNMAR_PARTNER	Unmarried partner
	FOSTER_CHILD	
	OTHER_NON_REL	Other nonrelative
	SAS_MISS_R	Unknown or invalid value
MARSTU	*EVER_MARRIED	A response class
	*NOT_NOW_MARRIED	A response class
	NOW_MARRIED	
	WIDOWED	
	DIVORCED	
	SEPARATED	
	NEVER_MARRIED	
	SAS_MISS_M	Unknown or invalid value

14. 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 if-then-else system is given in Table 4. The edit rule in this example is the "Universe" and "If" portions of the specification. They have to be converted into a computer code. When the edit rules are changed, the program must be rewritten. If the changes are substantial

Table 2: Valueset of the four response classes with NIM.

Response Class	Values	Notes
*EVER_MARRIED	NOW_MARRIED WIDOWED DIVORCED SEPARATED	
*NOT_NOW_MARRIED	WIDOWED DIVORCED SEPARATED NEVER_MARRIED	
*PARTNER	HUSBAND_WIFE UNMAR_PARTNER	 Unmarried partner
*RELATIVE	SON_DAUGHTER GRAND_CHILD IN_LAW ROOMER HOUSEMATE FOSTER_CHILD	 Roomer/boarder Housemate/roommate

Table 3: All Possible Values for sex, hhr, and ms with DISCRETE.

sex	household relationship (hhr)	marital status (ms)
`SEXU11, SEXU22, SEXU33`	`RELANU11, RELANU22, RELANU33`	`MARSTU11, MARSTU22, MARSTU33`
1 = Male 2 = Female 3 = Unknown	1 = Householder 2 = Husband/wife 3 = Son/daughter 4 = Brother/sister 5 = Father/mother 6 = Grandchild 7 = In-law 8 = Other relative 9 = Roomer/boarder 10 = Housemate/roommate 11 = Unmarried partner 12 = Foster child 13 = Other nonrelative 14 = Unknown	1 = Now married 2 = Widowed 3 = Divorced 4 = Separated 5 = Never married 6 = Unknown

and the programmers that wrote the previous code are not available, then the programs are typically completely rewritten.

Table 4: 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 TP(4); set allocation flag.

15. The NIM system uses decision logic tables (DLT) to store the edit rules. Unlike the if-then-else system, the DLTs are input to the NIM program. The changes of the edit rules only necessitate changes of the DLTs. 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 5. The first column of the Y's, N's, and spaces represents the edit rule described in Table 4. The last edit rule in Table 5 indicates that a householder's age has to be at least 15. 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 5: Decision Logic Table 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;
```

16. 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 4 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. The input format of the last edit rule in the DLT of Table 5 for the DISCRETE program is

```
Explicit edit # 40: 2 entering field(s)
  RELANU11      1 response(s):  1
  AGEU13        1 response(s):  1
```

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 listed in Table 9 in section .

17. To apply the different systems, we need two further refinements of the edit rules. The first are pre-edits that are implemented during the preprocessing of the data to prepare for the main edit/imputation. The pre-edits are much easier to develop because they typically involve a single field. They are described in the next section. The second is partitioning the age range [0,115] into sub-regions corresponding to edits. The age partitioning is described in section and is only needed for the DMB system.

IV. Pre-Edits

18. Some missing fields in a record can be logically derived from other non-missing fields. For example, a missing marital status can be filled in if we know the person is the spouse of the householder. This type of edit, also referred to as a logical edit, is called a *pre-edit*. 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.

19. 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 according to Table 6, 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*.

20. Many individual records in the ACS data have either age or date of birth missing or there exists inconsistency between the age and the date of birth. An edit rule to correct this type of error is usually called *within person edit rule*. The within person edit rules in this study for the age and date of birth

Table 6: Household Relationship Conversion Table.

hhr before the pre-edits	hhr after the pre-edits
1 Householder	3 Child
2 Spouse	7 In-law
3 Child	6 Grandchild
4 Sibling	3 Child
5 First Parent	1 Householder
5 Second Parent	2 Spouse
5 Third or Subsequent Parent	2 Spouse (see spouse pre-edits)
6 Grandchild	8 Other Relative
7 In-Law	8 Other Relative
8 Other Relative	8 Other Relative
9 Roomer or Boarder	9 Roomer or Boarder
10 Housemate or Roommate	13 Other Nonrelative
11 Unmarried Partner	13 Other Nonrelative
12 Foster Child	12 Foster Child
13 Other Nonrelative	13 Other Nonrelative
14 Unknown	14 Unknown

are used to impute the missing value if the other is not missing and valid. The marital status pre-edits are to make correction to the field of marital status if a person less than 15 is other than *never married*.

21. One of the important characteristics of NIM is that it requires a 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 7, which lists the percentage of failed records for different household sizes with and without pre-edits. Table 7 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 means that there is not enough donors to preserve the statistical properties of the survey data set.

22. Prior to running the DMB system, we need to put ages into a form that drastically reduces the amount of computation and facilitates age comparisons of individuals across age-generations within households. We also break up households having more than three individuals into subsets of three individuals that also facilitates the age comparisons. With the new representation, we generate a full set of implicit edits. As part of the production editing, we recombine the three-groups for households having more than three individuals.

23. We more fully describe the conversion of households into sets of three-person households. For convenience, we assume that there are at most three generations living in a household so that each household is converted into a three-person household. We assume that the householder and the spouse (or spouse-equivalent) if present are, respectively, the first and second members. The third member will be one

Table 7: Percentage Failed with NIM.

Household Size	*Total Households*	*without pre-edits*		*with pre-edits*	
		Failed Households	*Percentage Failed*	*Failed Households*	*Percentage Failed*
2	37120	10426	28.09	8580	23.11
3	16954	6786	40.03	4658	27.47
4	14258	6447	45.22	4785	33.56
5	6742	3429	50.86	2643	39.20
6	2129	1968	92.44	1189	55.85
7	719	685	95.27	438	60.92
8	319	308	96.55	207	64.89
9	150	145	96.67	96	64.00
Total	78391	30194	38.52	22596	28.83

of the others. For example, if a household has 4 persons: two parents and two children, then this four-person household is converted into two three-person households: the first household consists of the two parents and the first child and the second household the two parents and the second child. The conversion is consistent with the edits defined in the age comparison condition variables in Tables 8 and 9 of Section .

V. Age Decomposition of the DMB System

24. 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:

$$a_1x_1 + a_2x_2 + a_3x_3 > b, \tag{1}$$

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 (1) is true; 2 if false; and 3 if unknown. Table 8 lists the 41 temporary age comparison condition variables of inequality (1) for this study. For example, one of the 41 age comparison condition variables is $x_1 - x_2 > -12$ (Table 8 Inequality 14), where $a_1 = 1$, $a_2 = -1$, and $a_3 = 0$. If the first person's age is 35 and the second is 32, then the value of the variable of $x_1 - x_2 > -12$ is 1 because it is true that $35 - 32 > -12$. Another example is that the first person's age is less than or equal to 14: $x_1 \leq 14$, that is converted to the normalized form of $-x_1 > -15$ in (1) with $a_1 = -1$, $a_2 = a_3 = 0$, and $b = -15$ (Table 8 Inequality 1).

25. The 41 temporary Age comparison condition variables can be converted into six variables with the form (see Chen and Winkler (2002) for more details):

$$a_1x_1 + a_2x_2 + a_3x_3, \tag{2}$$

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 Fellgi-Holt model described in Section . Table 9 lists the six variables and their possible coded values for the intervals derived from the 41 inequalities

Table 8: The 41 Temporary Age Comparison Condition Variables.

Inequality ID	a_1	a_2	a_3	b	Inequality ID	a_1	a_2	a_3	b
1	−1	0	0	−15	22	1	0	−1	14
2	0	−1	0	−15	23	0	1	0	74
3	0	0	−1	−15	24	−1	1	0	−15
4	1	0	0	115	25	0	0	1	74
5	0	1	0	115	26	−1	0	1	−4
6	−1	1	0	−12	27	0	−1	1	−4
7	1	−1	0	49	28	0	1	−1	−15
8	−1	0	1	−12	29	1	0	−1	−15
9	1	0	−1	49	30	0	−1	0	−30
10	1	−1	0	34	31	0	0	−1	−30
11	−1	1	0	34	32	0	1	0	59
12	1	0	−1	34	33	0	0	1	59
13	−1	0	1	34	34	−1	0	1	−20
14	1	−1	0	−12	35	0	−1	1	−20
15	1	0	−1	−12	36	0	−1	1	−12
16	−1	1	0	−30	37	0	1	0	89
17	−1	0	1	−30	38	0	0	1	89
18	0	−1	0	−18	39	−1	0	0	−30
19	0	0	−1	−18	40	−1	0	1	−25
20	1	−1	0	0	41	0	−1	1	−25
21	0	1	−1	14					

for (1) listed in Table 8. The formulation significantly reduced the size of the set covering problem of the edit generation and the error localization.

VI. Existing If-Then-Else Rules Used by ACS

26. 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.

27. 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. In each section, there are several allocation matrices. For example, in the sex section, there is an allocation matrix for cases where "sex" is missing and "age" is not missing. If age is 61, the sex will be imputed to a value with 47% of male and 53% of female according to the matrix. 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 if-then-else 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 (see Example 7128560 in Section 10.1.1).

Table 9: The Six Variables Defined for Age Comparisons.

Variable Name	Form (1)	Coded Values	Variable Name	Form (1)	Coded Values
AGEU10	x_3	1 = [0, 14] 2 = [15, 17] 3 = [18, 29] 4 = [30, 59] 5 = [60, 74] 6 = [75, 89] 7 = [90, 999] 8 = unknown*	AGEU13	x_1	1 = [0, 14] 2 = [15, 29] 3 = [30, 115] 4 = [116, 999] 5 = unknown*
AGEU11	x_2	1 = [0, 14] 2 = [15, 17] 3 = [18, 29] 4 = [30, 59] 5 = [60, 74] 6 = [75, 89] 7 = [90, 115] 8 = [116, 999] 9 = unknown*	AGEU14	$x_1 - x_3$	1 = [−999, −35] 2 = [−34, −15] 3 = [−14, −12] 4 = [−11, 3] 5 = [4, 11] 6 = [12, 14] 7 = [15, 19] 8 = [20, 24] 9 = [25, 29] 10 = [30, 34] 11 = [35, 49] 12 = [50, 999] 13 = unknown*
AGEU12	$x_2 - x_3$	1 = [−999, −15] 2 = [−14, 3] 3 = [4, 11] 4 = [12, 14] 5 = [15, 19] 6 = [20, 24] 7 = [25, 999] 8 = unknown*	AGEU15	$x_1 - x_2$	1 = [−999, −35] 2 = [−34, −12] 3 = [−11, 0] 4 = [1, 11] 5 = [12, 14] 6 = [15, 29] 7 = [30, 34] 8 = [35, 49] 9 = [50, 999] 10 = unknown*
** if the age of at least one of the involved person(s) is unknown or invalid*					

28. The ACS ITE system is implemented with SAS programming language from SAS Institute Inc. 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. In this study, we did not go through the process described above, we simply extract the unedited and edited records from the data files that have been processed by the ACS staff.

VII. Bankier's Nearest-Neighbor Imputation Method

29. 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,

$$D_{fp} = D(\boldsymbol{a}_f, \boldsymbol{a}_p) = \sum_i \omega_i D_i(a_{fi}, a_{pi}) \tag{3}$$

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,

$$0 \leq D_i(a_{fi}, a_{pi}) \leq 1 \tag{4}$$

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 (3).

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

$$D_i(a_{fi}, a_{pi}) = \begin{cases} 1 & \text{if } |a_{fi} - a_{pi}| \geq m(a_{fi}) \\ 1 & \text{if the value of } a_{fi} \text{ is missing or invalid} \\ 1 & \text{if } a_{fi} \geq 15 \text{ and } a_{pi} < 15 \text{ (an adult} \\ & \text{to child conversion),} \\ 1 & \text{if } a_{fi} < 15 \text{ and } a_{pi} \geq 15 \text{ (a child} \\ & \text{to adult conversion)} \\ 1 - (1 - \frac{|a_{fi} - a_{pi}|}{m(a_{fi})})^r & \text{otherwise} \end{cases}$$

where r is a non-negative 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.

31. 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

$$D_{fpa} = \alpha D_{fa} + (1 - \alpha) D_{ap} \tag{5}$$

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.

32. 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

$$D_{fpa} \leq \gamma \min\{D_{fpa}\} \tag{6}$$

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.

33. There are two crucial advantages for a NIM system. The first is that (virtually) all of the imputed records satisfy all of the edits. The second is that it finds the best matching rules automatically. There is another important insight. By considering the set of fields in a donor record that differ from the edit-failing record, it is possible to efficiently fill-in (determine the subset of fields to change) a record. The potential value states are always two. Either leave the value in a field to its value in the original edit-failing record or change it to the value in the potential donor record. Although this does not always assure minimum change, it does assure that there is no combinatorial explosion of values that need to be substituted.

VIII. DISCRETE Edit and Model-Based Imputation System

VIII.1 DISCRETE Editing

34. 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, $a_i \in A_i$, $1 \leq i \leq n$, 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

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

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

35. 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:

$$\begin{aligned} &\text{Minimize} && \textstyle\sum_{j=1}^{n} c_j x_j \\ &\text{subject to} && \textstyle\sum_{j=1}^{n} a_{ij} x_j \geq 1, \quad i = 1, 2, \cdots, m \end{aligned} \tag{7}$$

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

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 (see Table 9 for the age comparison variables). We need to get $\bar{E}$ from a *complete* set of edits to obtain a meaningful solution to (7). A complete set of edits is the set of explicit (initially specified) edits and all essentially new implied edits derived from them.

36. If $\boldsymbol{x}$ is a prime cover solution to (7) 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 $J = \{j \mid 1 \leq j \leq m,\ f_k \text{ is an entering field of } E^j\}$. 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 $\bar{E}$.

37. 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. *Lemma* 1 gives a formulation on how to generate implicit edits.
Lemma 1 (Fellegi and Holt 1976): If E^r are edits $\forall\ r \in S$, where S is any index set,

$$E^r : \bigcap_{j=1}^{n} \boldsymbol{A}_j^r = F, \quad \forall \quad r \in S.$$

Then, for each i $(1 \leq i \leq n)$, the expression

$$E^* : \bigcap_{j=1}^{n} \boldsymbol{A}_j^* = F \tag{8}$$

is an implied edit, where

$$\boldsymbol{A}_j^* = \bigcap_{r \in S} \boldsymbol{A}_j^r \neq \emptyset \quad j = 1, \cdots, i-1, i+1, \cdots, n$$

$$\boldsymbol{A}_i^* = \bigcup_{r \in S} \boldsymbol{A}_i^r \neq \emptyset.$$

If all the sets A_i^r are proper subsets of A_i, i.e., $A_i^r \neq A_i$ (field i is an entering field of edit E^r) $\forall\ r \in S$, but $A_i^* = A_i$, then the implied edit (6) is called an *essentially new edit*. Field i, which has n_i possible values, is referred to as the *generating field* of the implied edit. The edits $E^r\ \forall\ r \in S$ from which the new implied edit E^* is derived are called *contributing edits*.

38. Therefore, in order to generate an essentially new implicit edit, we must have the following three conditions:

    (a) $A_j^* \neq \emptyset, \forall\ j,\ 1 \leq j \leq n$;

    (b) $A_i^r \neq A_i, \forall\ r \in S$, where $A_i^r \neq \emptyset$;

    (c) $A_i^* = A_i$.

    Conditions 2 and 3 indicates that the set $\{A_i^r \mid r \in S\}$ is a cover of A_i and are the foundations of the following set covering formulation in (9).

39. Let $\{E^r \mid r \in S\}$ be the set of the s edits with field i entering, then the set covering problem related to the generating field i is

$$\begin{aligned} &\text{Minimize} && \textstyle\sum_{r \in S} x_r \\ &\text{subject to} && \textstyle\sum_{r \in S} g_{rj}^i x_r \geq 1, \quad j = 1, 2, \cdots, n_i \end{aligned} \tag{9}$$

$$x_r = \begin{cases} 1, & \text{if field } E^r \text{ is in the cover;} \\ 0, & \text{otherwise,} \end{cases} \qquad r \in S$$

    where

$$g_{rj}^i = \begin{cases} 1, & \text{if } E^r \text{ contains the } j\text{th element in the field } i; \\ 0, & \text{otherwise,} \end{cases}$$

    is the jth element in field i of edit E^r ($r \in S$). If $\boldsymbol{x}$ is a prime cover solution to (7) and $K = \{r \mid x_r = 1\} \subset S$, then $\cup_{k \in K} A_i^k = A_i$. A prime cover solution is a nonredundant set of the edits whose ith components cover all possible values of the entering field, which is the generating field to yield an essentially new implicit edit.

40. 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 . The age comparison program is based on new age comparison variables given in Section . It is more fully described in Chen and Winkler (2002), which has a better performance than the one described in Chen, Winkler, and Hemmig (2000).

VIII.2 Model-Based Imputation

41. The imputation module MB of Discrete system is based on the general location model (Olkin and Tate 1971, Schafer 1997, Little and Rubin 2002). It produces single non-random item imputations based on the MLE of conditional probabilities derived from the model. This approach is similar to that of Thibaudeau (2002) in the sense that in both cases the imputations are derived directly from the MLE of conditional probabilities. These conditional probabilities are used to generate single imputations. The imputation methodology MB differs from the methodology of Thibaudeau (2002) in two ways. First, Thibaudeau used geospatial information available from the decennial census to derive random item imputation. MB does not use geospatial information to generate imputations. Geospatial information is not available for general surveys such as ACS. MB and NIM compensate for the unavailabilable geospatial information by aggregating households of the same size and then processing the edits and imputations. This means that MB and NIM use the intra-cluster correlation in terms of the profiles of the members of same-size households. Second, the methodology of Thibaudeau (2000) applies the EM algorithm for missing item imputation without regard to edit restraints. To avoid the additional complications associated with the edit constraints, MB does not use the EM algorithm. Parameter estimation for the general location model proceeds only from the records initially passing all the edits. This approach somewhat corresponds to NIM. NIM primarily only uses donors that pass all the edits to generate imputations. Both MB and NIM have the potential limitation that they are dependent on having a moderate number of donors. With more donors (i.e., edit-passing records), NIM is more likely to have a donor record that more closely resembles the edit-failing record; MB has more data for creating the model for item imputation. We expect that both MB and NIM imputation methodologies to perform well when there are high proportions of edit-passing records. With high proportions, MB and NIM have access to maximum information in order to impute items realistically.

42. The specifics of MB imputation are as follows. The general location model serves MB to impute 1. Categorical data (i.e. relationship sex, and marital status). 2. Continuous data (i.e. age). We illustrate the imputation of categorical data with an example. The imputation of continuous data is more straightforward and will be explained last. Table 10 shows the information reported by a specific household of four. While age is reported for each member of the household, sex is not reported for any member but the householder. Furthermore, the relationship of person 3 conflicts with his/her reported age, and so the Fellegi-Holt algorithm flags this relationship as an edit failure. The imputations for sex and relationship will be determined by a discriminant analysis of the age pattern of the household members relative to similar households. This analysis is based on the likelihood of the general location model. The likelihood is based on a mixture of normal kernels. Each kernel corresponds to a joint classification of the categorical data and to the corresponding mean household ages. For this analysis, the households are three-person sub-households made-up of the householder, the spouse, and any other household member. Information from the reported categorical variables of the members serves to identify possible categories of households that are eligible as the closest category. In our implementation, the only categories of households that are eligible for the status of closest category are those categories that are compatible with the household in Table 10 in terms of the relationship and sex of person 1 and 2, and the marital status of person 3. This leaves several possibilities for an imputed relationship and sex for person 3. The discriminant function indicates category that is closer.

The measure of distance is a trivariate Mahalanobis age distance in the square-root scale between the household in Table 10 and each household category in terms of its average trivariate square-root age. This distance function is weighted by the covariance matrix of the square root of the three ages, for each prospective category. For 1999 ACS data, the household category with the "closest" trivariate square-root age among the prospective categories dictates that the relationship of person 3 is "child" and the sex of person 3 is "male". For this category of the discriminant function is .542. The next two "closest" categories implicitly impute the relationship and sex of person 3 to be "child-female", and "other-relative-male", respectively. The corresponding values of the discriminant function are .436, and .007. The discriminant function is just a normalized version of the distance function described above. It can be interpreted as a probability because it adds-up to one over all the eligible household categories.

Table 10: Household A - Reported Information

Variable Name	Age	Relation with Householder	Marital Status	Sex
Householder	61	Self	Married	Male
Person 2	49	Spouse	Married	Missing
Person 3	21	Parent *(Flagged by Edit)*	Never Married	Missing
Person 4	86	In-Law	Widowed	Missing

43. MB automatically selects the most likely category of households and imputes the relationship-sex profile in Table 11 for person 3. Next, based on another matching rule involving the reported items, MB selects the most likely household category to impute the sex of person 2. It then selects yet another household category to impute the sex of person 4. The resulting item imputations in Table 11 are plausible. In particular, the sex of person 2 is female, as it should be. Given the age and the well-known longevity of females, it is not suprising the the imputed set of person 4 is female.

Table 11: Household A - Reported and Imputed Information

Variable Name	Age	Relation with Householder	Marital Status	Sex
Householder	61	Self	Married	Male
Person 2	49	Spouse	Married	*Female*
Person 3	21	*Child*	Never Married	*Male*
Person 4	86	In-Law	Widowed	*Female*

IX. Statistical Comparisons

44. 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 four 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. For example, Table 12 lists the frequency distributions of the marital status for the 4-person edit-passing households and imputed households by NIM.

Table 12: The Frequency Distributions of the Marital Status for the 4-person Households.

	edit-passing households		imputed households by NIM	
value (i)	frequency (r_i)	proportion (x_i)	frequency (s_i)	proportion (y_i)
1. married	14721	0.3896	1250	0.3853
2. widowed	573	0.0152	52	0.0160
3. divorced	1414	0.0374	107	0.0330
4. separated	507	0.0134	37	0.0114
5. never married	20569	0.5444	1798	0.5543
total	37784	1.0000	3244	1.0000

45. 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:

$$\sum_{i=1}^{n}(x_i - y_i)^2, \tag{10}$$

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. Table 3 lists the categories for sex, household relationship, and marital status, with "Unknown" category excluded. 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.

46. In the comparisons among the three systems, a small value of the sum of squared deviations of 10 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 of equation (10). Table 13 lists the values of equation (10) for the three systems by variables and household sizes. The column of "sum" is the sum of the values from columns "3-person" to "9-person" representing the aggregate measurement of each of the univariate and bivariate frequency distributions. From Table 13, it is clear that NIM outperforms the existing If-Then-Else (ITE) and the DMB (DISCRETE Model-Based) systems in term of the measurement of the sum of squared deviation of equation (10).

X. Comparisons of Imputed Results

47. In this section, we discuss the comparisons of the imputed results of the edit failing households from the three systems. Comparing the numbers in the column "*Total Households Imputed* " of Table 14

Table 13: Comparisons of Sum of Squared Deviations.

variable	system	3-person	4-person	5-person	6-person	7-person	8-person	9-person	sum
sex	ITE	0.0012	0.0014	0.0002	0.0014	0.0001	0.0014	0.0113	0.0170
	NIM	0.0007	0.0011	0.0000	0.0018	0.0002	0.0003	0.0095	0.0136
	DMB	0.0035	0.0052	0.0007	0.0035	0.0013	0.0039	0.0006	0.0187
ms	ITE	0.0021	0.0019	0.0019	0.0128	0.0175	0.0171	0.0188	0.0721
	NIM	0.0009	0.0001	0.0000	0.0012	0.0005	0.0010	0.0046	0.0083
	DMB	0.0005	0.0002	0.0003	0.0022	0.0008	0.0004	0.0017	0.0061
age	ITE	0.0027	0.0030	0.0043	0.0054	0.0100	0.0041	0.0175	0.0470
	NIM	0.0011	0.0015	0.0017	0.0005	0.0009	0.0018	0.0046	0.0121
	DMB	0.0022	0.0021	0.0033	0.0028	0.0165	0.0058	0.0106	0.0433
hhr	ITE	0.0169	0.0216	0.0193	0.0123	0.0102	0.0027	0.0120	0.0950
	NIM	0.0021	0.0025	0.0029	0.0006	0.0014	0.0005	0.0038	0.0138
	DMB	0.0067	0.0054	0.0066	0.0019	0.0052	0.0066	0.0032	0.0356
sex-ms	ITE	0.0016	0.0015	0.0010	0.0068	0.0088	0.0090	0.0183	0.0470
	NIM	0.0008	0.0006	0.0001	0.0014	0.0004	0.0006	0.0110	0.0149
	DMB	0.0021	0.0036	0.0011	0.0029	0.0012	0.0037	0.0012	0.0158
sex-age	ITE	0.0017	0.0017	0.0024	0.0030	0.0055	0.0035	0.0133	0.0311
	NIM	0.0007	0.0009	0.0009	0.0006	0.0010	0.0021	0.0066	0.0128
	DMB	0.0018	0.0022	0.0019	0.0019	0.0092	0.0081	0.0100	0.0351
sex-hhr	ITE	0.0096	0.0113	0.0101	0.0068	0.0060	0.0024	0.0125	0.0587
	NIM	0.0023	0.0019	0.0015	0.0013	0.0024	0.0009	0.0099	0.0202
	DMB	0.0050	0.0055	0.0044	0.0034	0.0061	0.0071	0.0037	0.0352
ms-age	ITE	0.0032	0.0043	0.0048	0.0049	0.0091	0.0042	0.0148	0.0453
	NIM	0.0021	0.0019	0.0019	0.0005	0.0008	0.0018	0.0048	0.0138
	DMB	0.0029	0.0026	0.0035	0.0029	0.0164	0.0055	0.0102	0.0440
ms-hhr	ITE	0.0179	0.0226	0.0212	0.0118	0.0098	0.0050	0.0171	0.1054
	NIM	0.0034	0.0027	0.0034	0.0011	0.0022	0.0019	0.0041	0.0188
	DMB	0.0105	0.0069	0.0079	0.0024	0.0061	0.0078	0.0044	0.0460
age-hhr	ITE	0.0051	0.0070	0.0079	0.0043	0.0062	0.0024	0.0123	0.0452
	NIM	0.0015	0.0018	0.0024	0.0007	0.0016	0.0021	0.0047	0.0148
	DMB	0.0028	0.0027	0.0035	0.0019	0.0128	0.0060	0.0127	0.0424

and the column "*Failed Households* " with pre-edits of Table 7, we found thateleven of the 4785 failed 4-person households were not recorded in the NIM output file. That was because the number of failed households was over 100 before the passed ones reaching 100 in some imputation groups. There were 161 of the failed 6-person households that were not recorded. We believe this limitation would be eliminated in the future releases of NIM.

48. According to Table 14, the total number of households imputed for this study is 13844.There are 10,689 households, or 77.2%, that have exactly the same imputed results with the If-Then-Else rules and NIM. The other 3,155 households, or 22.8%, have at least one imputed values disagreed. Table 15 lists the numbers of imputed households that have "ndif" imputed values disagreed between ITE and NIM. Tables 16 and 17 are for between ITE and DMB and between NIM and DMB, respectively.

49. We also compare the imputed results from the three systems that still failed at least one of the edits specified in the edit table of the DISCRETE edit system. Table 18 lists the percentage of the households failed at least one edit after the imputations. The results indicate that the ITE system provides the best imputation in term of this measurement of still-failing-edit after imputation.

Table 14: Agreed and Disagreed of the Three Systems.

Household Size	*Total Households Imputed*	*ITE vs. NIM*		*ITE vs. DMB*		*NIM vs. DMB*	
		Agreed	*Disagreed*	*Agreed*	*Disagreed*	*Agreed*	*Disagreed*
3	4658	3564	1094	3627	1031	3632	1026
4	4774	3958	816	3961	813	4038	736
5	2643	2007	636	2056	587	2091	552
6	1028	720	308	736	292	728	300
7	438	269	169	299	139	276	162
8	207	117	90	144	63	121	86
9	96	54	42	63	33	56	40
Total	13844	10689	3155	10886	2958	10942	2902

Table 15: Numbers of Disagreed Imputed Values between ITE and NIM.

HH size	Number of Fields Disagreed (ndif)																		Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
3	466	369	176	114	26	19	15	2	0	0	0	0	–	–	–	–	–	–	1187
4	353	236	143	108	21	12	12	4	2	3	2	0	0	0	0	0	–	–	896
5	237	171	117	112	27	20	11	14	9	5	2	2	1	0	1	0	0	0	729
6	50	77	102	81	20	6	11	6	3	1	2	0	1	1	1	0	0	0	362
7	18	28	34	28	18	23	33	14	3	3	7	3	1	0	1	0	0	0	214
8	12	24	20	14	8	13	14	9	7	10	8	2	0	0	0	0	0	0	141
9	7	7	6	8	2	1	1	2	0	1	3	3	1	4	0	2	1	1	50
Total	1143	912	598	465	122	94	97	51	24	23	24	10	4	5	3	2	1	1	3579

Table 16: Numbers of Disagreed Imputed Values between ITE and DMB.

HH size	Number of Fields Disagreed (ndif)																		Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
3	516	272	119	93	22	7	2	0	0	0	0	0	–	–	–	–	–	–	1031
4	384	195	129	73	13	9	5	2	1	2	0	0	0	0	0	0	–	–	813
5	268	141	66	59	22	17	5	1	3	2	2	0	0	1	0	0	0	0	587
6	56	67	78	55	17	5	5	2	3	1	1	1	1	0	0	0	0	0	292
7	21	19	12	14	16	18	23	9	2	2	1	0	1	0	0	1	0	0	139
8	14	8	6	2	1	3	4	2	8	8	4	1	1	0	1	0	0	0	63
9	7	5	1	0	3	1	0	0	1	0	2	4	5	2	1	0	1	0	33
Total	1266	707	411	296	94	60	44	16	18	15	10	6	8	3	2	1	1	0	2958

Table 17: Numbers of Disagreed Imputed Values between NIM and DMB.

HH	Number of Fields Disagreed (ndif)																		
size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
3	544	299	171	53	22	20	2	1	0	0	0	0	–	–	–	–	–	–	1112
4	388	208	118	57	17	9	8	5	3	1	0	0	0	0	0	0	–	–	814
5	234	158	109	66	27	15	20	11	6	1	0	0	0	0	0	0	0	0	647
6	110	100	59	35	19	10	9	8	1	2	1	0	0	0	0	0	0	0	354
7	18	41	50	33	18	13	18	9	3	4	1	0	0	0	0	0	0	0	208
8	12	20	20	19	15	13	16	9	6	2	0	3	1	0	0	1	0	0	137
9	4	8	9	6	4	2	4	3	2	1	0	2	2	1	0	0	0	0	48
Total	1310	834	536	269	122	82	77	46	21	11	2	5	3	1	0	1	0	0	3320

Table 18: Percentage(%) of Households Failed after Imputations.

Household Size	*ITE*	*NIM*	*DMB*
3	0.22	0.45	0.61
4	0.34	0.51	0.62
5	0.48	1.22	0.96
6	1.18	1.64	1.90
7	1.50	1.24	2.40
8	1.42	2.30	2.13
9	2.10	2.02	2.80
Total	0.39	0.78	0.79

50. In the following subsections, we will examine several imputed households that are still problematic after the imputations. We intend to provide some information on what can be possibly improved to have "better" or "reasonable" imputation results. These information might be particularly useful for the If-Then-Else rules because of the sequential editing nature of the system.

X.1 Problems with If-Then-Else

51. When a married *child* or *parent* has similar age of the married householder and the spouse is missing in the household, the If-Then-Else system calls this person *other relative*, and NIM and DMB call him/her *spouse*, for example Table 19
52. When a married *unmarried partner* has similar age of the married householder and the spouse is missing in the household, the If-Then-Else system calls this person *roomer/boarder*, *brother/sister*, or *other relative*; NIM calls him/her *spouse*, Table 20 is an example. In this example, DMB keeps the relationship of the fifth person and changes the marital status of the first and fifth persons. Does it violate the criterion of changing the minimum number of fields? The answer is no because of the the measure of confidence in field j, c_j, given in (7). The measure of confidence in the field of household relationship used in this study was 5.20 and that in the field of marital status 2.10, which makes the optimal value of 4.20 in (7) from the two fields of marital status. This example is different from Example 8316669, that involves the age comparison between the 56-year-old householder and his

Table 19: Example 8316669.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	56	householder	married	male	56	householder	married
2	female	16	daughter	n. married	female	16	daughter	n. married
3	male	14	son	n. married	male	14	son	n. married
4	female	53	daughter	married	female	53	spouse	married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	56	householder	married	male	56	householder	married
2	female	16	daughter	n. married	female	16	daughter	n. married
3	male	14	son	n. married	male	14	son	n. married
4	female	53	other relative	married	female	53	spouse	married

53-year-old "daughter".

Table 20: Example 6678946.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	41	householder	married	male	41	householder	married
2	female	18	mother	n. married	female	18	daughter	n. married
3	male	19	son	n. married	male	19	son	n. married
4	female	16	other relative	n. married	female	16	other relative	n. married
5	female	38	unm. partner	married	female	38	spouse	married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	41	householder	married	male	41	householder	widowed
2	female	18	daughter	n. married	female	18	mother	n. married
3	male	19	son	n. married	male	19	son	n. married
4	female	16	other relative	n. married	female	16	other relative	n. married
5	female	38	rommer/boarder	married	female	38	unm. partner	divorced

53. Table 21 is a typical example of ineffective sequential edit system, such as the If-Then-Else system. After imputing a value for the second person's *marital status*, the If-Then-Else system made an unnecessary change of the third person's age to 24, that fails the edit of the householder's age must be at least 15 years older than a child.

54. Unnecessary change of the fourth person's *SEX* by the If-Then-Else rules is given in Table 22.

55. Unnecessary change of the fifth person's *AGE* by the If-Then-Else rules (Table 23).

56. In this example (Table 24), the If-Then-Else rules gave a change of the third person's relationship to *son* that still fails the edit of a child having to be at least 15 years younger than the parent. This is another example of ineffective sequential edit system of the If-Then-Else rules. NIM seems making the minimum change of the household, in which the third person's relationship is changed from *unmarried partner* to *spouse*. DMB makes the changes according to the minimization problem given in (7).

Table 21: Example 7128560.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	female	36	householder	married	female	36	householder	married
2	male	37	spouse	unknown	male	37	spouse	married
3	female	12	daughter	n. married	female	12	daughter	n. married
4	male	10	son	n. married	male	10	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	female	36	householder	married	female	36	householder	married
2	male	37	spouse	married	male	37	spouse	married
3	female	24	daughter	n. married	female	12	daughter	n. married
4	male	10	son	n. married	male	10	son	n. married

Table 22: Example 7396046.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	46	householder	married	male	46	householder	married
2	female	35	spouse	married	female	35	spouse	married
3	male	16	father	n. married	male	16	son	n. married
4	female	6	mother	n. married	female	6	daughter	n. married
5	male	5	son	n. married	male	5	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	46	householder	married	male	46	householder	married
2	female	35	spouse	married	female	35	spouse	married
3	male	16	son	n. married	male	16	son	n. married
4	male	6	son	n. married	female	6	daughter	n. married
5	male	5	son	n. married	male	5	son	n. married

Table 23: Example 6817928.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	35	householder	married	male	35	householder	married
2	female	29	spouse	unknown	female	29	spouse	married
3	male	11	son	n. married	male	11	son	n. married
4	female	0	daughter	n. married	female	0	daughter	n. married
5	female	17	in-law	n. married	female	17	in-law	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	35	householder	married	male	35	householder	married
2	female	29	spouse	married	female	29	spouse	married
3	male	11	son	n. married	male	11	son	n. married
4	female	0	daughter	n. married	female	0	daughter	n. married
5	female	21	in-law	n. married	female	17	in-law	n. married

Table 24: Example 5970235.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	female	35	householder	married	female	35	householder	married
2	female	15	daughter	married	female	15	daughter	married
3	male	21	unm. partner	married	male	21	spouse	married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	female	35	householder	married	female	35	householder	separated
2	female	15	daughter	married	female	15	daughter	married
3	male	21	son	married	male	21	unm. partner	n. married

X.2 Problems with NIM

57. Table 25 provides 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*. The If-Then-Else rules also change the second person's relationship to *spouse*.

Table 25: Example 5839240.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	33	householder	married	male	33	householder	married
2	female	29	o. nonrelative	married	female	29	o. nonrelative	married
3	male	2	unknown	n. married	male	2	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	33	householder	married	male	33	householder	married
2	female	29	spouse	married	female	29	o. nonrelative	married
3	male	2	son	n. married	male	2	son	n. married

58. Table 26 is an 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 If-Then-Else rules provide.

Table 26: Example 7062208.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	41	householder	married	male	41	householder	married
2	female	43	spouse	married	female	43	spouse	married
3	female	9	unknown	n. married	female	9	other relative	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	41	householder	married	male	41	householder	married
2	female	43	spouse	married	female	43	spouse	married
3	female	9	daughter	n. married	female	9	daughter	n. married

59. Unnecessary change of the first person's *AGE* by NIM (Table 27).

Table 27: Example 7399300.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	46	householder	married	male	47	householder	married
2	female	41	spouse	married	female	41	spouse	married
3	male	13	son	n. married	male	13	son	n. married
4	male	11	brother	n. married	male	11	son	n. married
5	male	7	brother	n. married	male	7	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	46	householder	married	male	46	householder	married
2	female	41	spouse	married	female	41	spouse	married
3	male	13	son	n. married	male	13	son	n. married
4	male	11	son	n. married	male	11	son	n. married
5	male	7	son	n. married	male	7	son	n. married

X.3 Problems with DMB

60. The major problem with the DISCRETE Model-Based imputation is that some of the imputed households still fail some of the edits. Examples are given below:

61. The following example (Table 28) indicates that the 52-year-old householder has a "daughter" of 50 years old after the DISCRETE Model-Based imputation.

Table 28: Example 5173782.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	52	householder	married	male	52	householder	married
2	female	50	unknown	unknown	female	6	daughter	n. married
3	male	12	son	n. married	male	12	son	n. married
4	female	41	spouse	unknown	female	41	spouse	married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	52	householder	married	male	52	householder	married
2	female	50	spouse	married	female	50	daughter	divorced
3	male	12	son	n. married	male	12	son	n. married
4	female	41	sister	married	female	41	sister	n. married

62. In the following household (Table 29), the 5-year-old daughter should not be married.

63. In the following household (Table 30, the 69-year-old householder should not have a father with age of 63.

X.4 Other Problems

64. Table 31 is an example that the three systems did not provide a *reasonable* imputed value for the marital status of the fourth person. A value of *married* seems a better choice.

Table 29: Example 6225023.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	female	66	householder	widowed	female	36	householder	married
2	male	31	spouse	married	male	31	spouse	married
3	female	31	daughter	married	female	10	daughter	n. married
4	female	2	daughter	n. married	female	2	daughter	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	female	66	householder	married	female	32	householder	married
2	male	31	spouse	married	male	31	spouse	married
3	female	31	daughter	married	female	5	daughter	married
4	female	2	grandchild	n. married	female	2	daughter	n. married

Table 30: Example 7491862.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	female	45	householder	separated	female	45	householder	separated
2	male	63	son	n. married	male	11	son	n. married
3	female	43	daughter	n. married	female	15	daughter	n. married
4	female	41	daughter	n. married	female	4	daughter	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	female	45	householder	separated	female	69	householder	separated
2	male	5	son	n. married	male	63	father	n. married
3	female	8	daughter	n. married	female	43	daughter	n. married
4	female	13	daughter	n. married	female	41	o. nonrel.	n. married

Table 31: Example 6668810.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	47	householder	n. married	male	47	householder	n. married
2	male	33	brother	n. married	male	33	brother	n. married
3	male	79	father	married	male	79	father	married
4	female	72	mother	unknown	female	72	mother	widowed
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	47	householder	n. married	male	47	householder	n. married
2	male	33	brother	n. married	male	33	brother	n. married
3	male	79	father	married	male	79	father	married
4	female	72	mother	divorced	female	72	mother	widowed

65. Table 32 is an example that both the If-Then-Else system and NIM made an unnecessary change of the third person's relationship.

Table 32: Example 5903670.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	female	45	householder	married	female	45	householder	married
2	male	53	spouse	unknown	male	53	spouse	married
3	male	18	foster child	n. married	male	18	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	female	45	householder	married	female	45	householder	married
2	male	53	spouse	married	male	53	spouse	unknown
3	male	18	other nonrelative	n. married	male	18	foster child	n. married

66. The marital status of the first and second persons is anything but *married* because the second person's relationship is *unmarried partner* (Table 33).

Table 33: Example 6459254.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	female	41	householder	unknown	female	41	householder	married
2	male	39	unm. partner	unknown	male	39	spouse	married
3	male	66	father	widowed	male	66	father	widowed
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	female	41	householder	married	female	41	householder	divorced
2	male	39	brother	married	male	39	unm. partner	n. married
3	male	66	father	widowed	male	66	father	widowed

67. Table 34 shows the unnecessary change of a *foster child* to *roomer/boarder* by If-Then-Else rules and to *child* by NIM.

Table 34: Example 6599274.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	38	householder	n. married	male	38	householder	n. married
2	male	2	son	n. married	male	2	son	n. married
3	unknown	20	foster child	n. married	male	20	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	38	householder	n. married	male	38	householder	n. married
2	male	2	son	n. married	male	2	son	n. married
3	male	20	roomer/boarder	n. married	female	20	foster child	n. married

68. Table 35 shows the imputed household by ITE and NIM still fails the edit of *a child must be at least 15 years younger.*

Table 35: Example 6518122.

	sex	*age*	*household relationship*	*marital status*	*sex*	*age*	*household relationship*	*marital status*
ID	*The household after pre-edits*				*Imputation by NIM*			
1	male	27	householder	married	male	27	householder	married
2	female	27	spouse	unknown	female	27	spouse	married
3	female	16	daughter	n. married	female	15	daughter	n. married
4	male	14	son	n. married	male	14	son	n. married
5	male	12	son	n. married	male	12	son	n. married
6	male	11	son	n. married	male	11	son	n. married
7	male	9	son	n. married	male	9	son	n. married
8	male	7	son	n. married	male	7	son	n. married
9	male	1	son	n. married	male	1	son	n. married
ID	*Imputation by If-Then-Else*				*Imputation by DMB*			
1	male	27	householder	married	male	27	householder	married
2	female	27	spouse	married	female	27	spouse	married
3	female	3	daughter	n. married	female	3	daughter	n. married
4	male	14	son	n. married	male	3	son	n. married
5	male	12	son	n. married	male	12	son	n. married
6	male	11	son	n. married	male	11	son	n. married
7	male	9	son	n. married	male	9	son	n. married
8	male	7	son	n. married	male	7	son	n. married
9	male	1	son	n. married	male	1	son	n. married

XI. Discussion and Summary

69. The results of this study indicate that NIM and DISCRETE always identify the same edit-passing and edit-failing household records. 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 three 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 outperforms the existing If-Then-Else system and the DISCRETE Model-Based imputation. An 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.

70. The comparison study in this paper is based on the assumption that the If-Then-Else Rules, the Decision Logic Table of NIM, and the edit table of DISCRETE are consistent or "identical" in term of the edit specifications. We don't have a procedure or methodology to prove that they are consistent

or "identical". Fortunately, we were able to identify the same edit-passing and edit-failing household records using the DLT of NIM and the edit table of DISCRETE.

The three systems still fail to correct some of the households that initially failed some of the edits. The DISCRETE edit system is an exact method. For an edit-failing record, it identifies the minimum number of fields to change as well as the field values to change to so that the imputed record would pass all of the edits. To improve the DISCRETE Model-Based imputation, we would propose a new research topic about the model-based imputation conditional on the identified field values to change to. For NIM, it should have a capability to impute a record from more than a donor. Default donors should also be given if there are not enough donors from the households with the same size. The default donors should be generated from the edit-passing households with different size to meet the requirement of preserving the distribution of the data. For the IF-Then-Else rules, it should overcome the nature of the sequential edit and to improve the programming tedchniques employeed to implement the If-Then-Else rules. For example, the programming should include a routine to analyze the edit-failing household based on the edit rules before an imputation action is taken.

REFERENCES

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

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

Chen, B. (1998), "Set Covering Algorithm in Edit Generation," *American Statistical Association, Proceedings of the Section on Statistical Computing,* 91–96.

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.

De Waal, T. (2000), "New Developments in Automatic Edit and Imputation at Statistics Netherlands," U.N. Economic Commission for Europe Work Session on Statistical Data Editing, Cardiff, UK, October 2000

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

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data, 2nd Edition*, John Wiley: New York.

Olkin, I. and Tate, R. F. (1961), "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, **32**, 448-465.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Schiopu-Kratina, I. and Kovar, J.G. (1989), "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Statistics Canada, Methodology Branch Working Paper BSMD 89-001E.

Thibaudeau, Y. (2002), "Model Explicit Item Imputation for Demographic Categories," *Survey Methodology,* **28**:135–143.

Winkler, W.E. (1995), "Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods, 467–472.*

Winkler, W.E. (1997), "Set-Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods,* 564–569.

Winkler, W.E. and Chen, B. (2002), "Extending the Fellegi-Holt Model of Statistical Data Editing," Research Report Statistics 2002-02, Statistical Research Division, Bureau of the Census, Washington, D.C.