BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR2000/03

# Using the DISCRETE Edit System for ACS Surveys

Bor-Chung Chen, William E. Winkler, Robert J Hemmig
Statistical Research Division
Methodology and Standards Directorate
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: 8/14/2000

# Using the DISCRETE Edit System for ACS Surveys*

**Bor-Chung Chen, William E. Winkler,** and **Robert J. Hemmig**
**Bureau of the Census**
**Washington, DC 20233-9100**

August 14, 2000

**Abstract**

This paper describes the application of the DISCRETE edit system to the American Community Surveys (ACS) for the questions of sex, age, householder relationship, marital status, and race. In order to compare and edit the ages of the persons in the same household, each household with more than 3 persons is converted into a three-person household. We will discuss how the edit system is used to incorporate the age comparison into an edit table, which is the input of the system. The DISCRETE edit system is based on the Fellegi and Holt model [1976] of editing. Advantages of using the DISCRETE edit system include that the logical consistency of the edit system can be performed before the real production begins and that an edit table is used as an input file instead of the if-then-else rules.

KEY WORDS: Explicit Edits, Redundant Covers, Subcovers, Integer Programming, Optimization

## 1. Introduction

The American Community Survey (ACS), as part of the decennial program, is an on-going survey conducted by the U.S. Census Bureau that provides accurate and up-to-date profiles of America's Communities every year. The goals of the ACS are to (1) provide federal, state, and local governments an information base for the administration and evaluation of government programs; (2) improve the 2010 Census; (3) provide data users with timely demographic, housing, social, and economic data updated every year that can be compared across states, communities, and population groups. Full implementation of the

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

survey would begin in 2003 in every county of the United States. The survey would include three million households. Data are collected by mail and Census Bureau staff will follow-up those who do not respond. The edit methodology described in this paper together with a separate imputation method is a pilot study for the full implemenation of the survey. Therefore, this study only deals with the questions of sex (sex), age (age), householder relationship (hhr), marital status (ms), and race (race).

The information gathered in any survey, including the American Community Survey, may contain inconsistent or incorrect data. These erroneous data need to be revised prior to data tabulations and retrieval. The revisions of the erroneous data should not affect the statistical inferences of the data. One of the important steps of this systematic revision process is computer editing. Fellegi and Holt [1976] provided the underlying basis of developing a computer editing system. An edit-generation algorithm, called the EGE algorithm, for the DISCRETE edit system was described in Winkler [1997]. The EGE algorithm is a much faster alternative to Algorithm 1, called the GKL algorithm, of Garfinkel, Kunnathur, and Liepins [1986].

The objective of the DISCRETE edit system is to find the minimum number of fields to change in a record if the record fails a set of edits. This can be done in two stages. The first stage is to generate a complete set of edits. If no new edits can logically be derived from the explicit edits and known implied edits, the set of *all* edits (explicit and implied) having this property is called a *complete set of edits.* Fellegi and Holt [1976] showed that the edit generation process can check the edit system for logical consistency in which no contradictory edits exist. The set of failed edits of a record has to come from the complete set of edits for the imputed record to pass all the edits. The second stage is to find the minimum number of fields to change using the technique of the integer linear programming or the set covering problem. Because most of the information needed for error localization (finding the minimum number of fields to impute) comes from the edit generation, the production editing software is exceedingly fast.

We propose an editing methodology for the ACS study. The methodology consists of four programs:

1. the age comparison program: this program produces the explicit edits from the age comparison condition variables; the explicit edits produced are contradictions between the condition variables and are part of the input to the DISCRETE edit generation program.

2. the DISCRETE edit generation program: this program uses the Fellegi-

Holt model to generate a complete set of edits, removes redundant edits, and checks inconsistent edits; this step of edit generation can be completed before the survey data are available for production.

3. the pre-edit program: this program identifies the householder and spouse if present; it converts all households into at most three-person households; it also performs age, race, householder relationship, and marital status pre-edits; it generates date of birth if missing, and performs consistency checking between age and date of birth.

4. the error localization program: this program finds the minimum number of fields to impute if a record fails a set of edits; it uses the integer programming and the set covering algorithm to obtain the optional solution(s).

The edit methodology is straightforward to learn. It can be much easier to maintain and apply in a variety of situations because edit rules are contained in tables. The mathematical software routines do not need to be changed.

## 2. Explicit Edits

An edit is a record or set of records in which certain combinations of values or code values in different fields (corresponding to the different questions on a questionnaire) are unacceptable or not allowed.

Suppose that a record, $\boldsymbol{a}= (a_1, a_2, \ldots, a_n)$, has $n$ fields. $a_i \in A_i$ for each $i$ $1 \leq i \leq n$, where $A_i$ is the set of possible values or code values which may be recorded in Field $i$. $|A_i| = n_i$. If $a_i \in A_i^o \subseteq A_i$, we also say

$$\boldsymbol{a} \in \boldsymbol{E}_o = A_1^o \times A_2^o \times \cdots \times A_n^o.$$

The code space is $A_1 \times A_2 \times \ldots \times A_n = \boldsymbol{A}$. If each data point in $\boldsymbol{E}_o$ is an unacceptable code combination and $\exists$ at least one $i$, $1 \leq i \leq n$, $\ni$ $A_i^o \neq A_i$ and $\forall$ $j$, $1 \leq j \leq n$, $A_j^o \neq \emptyset$, $\boldsymbol{E}_o$ becomes an edit and is declared a SET of unacceptable code combinations. Record $\boldsymbol{a}$ is said to *fail* the edit specified by $\boldsymbol{E}_o$. If an explicit edit fails, then at least one value in an entering field must be changed. The record with the changed value in the entering field will no longer fail the explicit edit.

The expression of edits

$$A_1^o \times A_2^o \times \cdots \times A_n^o = F$$

is referred to as the *normal form of edits*. The set of edit rules in the normal form, as specified by the subject matter experts, is referred to as *explicit edits*. For example, suppose that a questionnaire contains three fields.

| *Field* | *Possible Codes* |
|---|---|
| Age | 0-14, 15+ |
| Marital Status | Single, Married, Divorced, Widowed, Separated |
| Relationship to Householder | Householder, Spouse of Householder, Other |

Ever Married = {Married, Divorced, Widowed, Separated}
Not Now Married = {Single, Divorced, Widowed, Separated}

There are two edits:

Edit 1. (Age $< 15$) and (MS = Ever Married) = F
Edit 2. (MS = Not Now Married) and (HHR = Spouse) = F

Let $A_1 = \{1, 2\}$, $|A_1| = 2$, where 1 = 0–14, 2 = 15+; $A_2 = \{1, 2, 3, 4, 5\}$, $|A_2| = 5$, where 1 = single, 2 = married, 3 = divorced, 4 = widowed, 5 = separated; $A_3 = \{1, 2, 3\}$, $|A_3| = 3$, where 1 = householder, 2 = spouse, 3 = other; then

$$\text{Edit 1: } \boldsymbol{E}_1 = A_1^1 \times A_2^1 \times A_3^1 = \{1\} \times \{2, 3, 4, 5\} \times A_3$$
$$\text{Edit 2: } \boldsymbol{E}_2 = A_1^2 \times A_2^2 \times A_3^2 = A_1 \times \{1, 3, 4\} \times \{2\}$$

To identify the explicit edits 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. Section 5 will describe how the households with more than 3 members are handled to meet this assumption. Therefore, the first nine of the 31 fields (or variables) identified are sex, householder relationship, and marital status for the three members in the household:1sex (meaning the first person's sex, which has the field ID or variable ID of var1), 1hhr (var2), 1ms (var3), 2sex (var4), 2hhr (var5), 2ms (var6), 3sex (var7), 3hhr (var8), and 3ms (var9). Table 1 lists the variable names and their possible code values. The other 22 fields are for the age comparison condition variables. Section 3 will give a more detailed description of the age comparisons.

A total of 163 explicit edits has been identified for this study. Twenty-nine of them directly came from the 1997 ACS Edit and Allocation Specifications. For example, in the 1997 ACS Edit and Allocation Specifications, an if-then-else rule indicates that

| | |
|---|---|
| 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. |

This rule is translated into the normal form of the edit:

$$A_1 \times \cdots \times A_4 \times \{2\} \times \{2, 3, 4, 5, 6\} \times A_7 \times \cdots \times A_{31} = F$$

with $A_5^o = \{2\}$ (var5) and $A_6^o = \{2,3,4,5,6\}$ (var6). Fields 5 and 6 are called *entering fields* of the edit because $A_5^o \neq A_5$ and $A_6^o \neq A_6$. The edit places restrictions on the values that fields 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 #100:  2 entering field(s)
  VAR5              1 response(s):  2
  VAR6              5 response(s):  2  3  4  5  6
```

The other 134 explicit edits are the results of contradictions specified in the age comparison conditions as described in Section 3.

**Table 1. All Possible Values for sex, hhr, and ms.**

| sex | householder relationship (hhr) | marital status (ms) |
|---|---|---|
| var1, var4, var7 | var2, var5, var8 | var3, var6, var9 |
| 1 = male<br>2 = female<br>3 = unknown | 1 = householder<br>2 = spouse<br>3 = child(natural/step)<br>4 = sibling<br>5 = parent<br>6 = grandchild<br>7 = in-law<br>8 = other relative<br>9 = roomer or boarder<br>10 = housemate or roommate<br>11 = unmarried partner<br>12 = foster child<br>13 = other nonrelative<br>14 = unknown | 1 = now married<br>2 = widowed<br>3 = divorced<br>4 = separated<br>5 = never married<br>6 = unknown |

## 3. Age Comparison

Each person in a three-person household has 9 fields for this study: sex, age, householder relationship (hhr), marital status (ms), race, detailed write-in race (p6cd), birth day (p2d), birth month (p2m), and birth year (p2y). The fields of sex, hhr, and ms are taken directly into the Fellgi-Holt model as described in Section 2. The other fields, except age, are for the pre-edits and the field of age is for the age comparison before the Fellgi-Holt model is used.

In the age comparison, each time when a new age restriction appears in one of the if-then-else rules in the 1997 ACS Edit and Allocation Specifications, a new age comparison condition variable is defined.An 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 $i$th 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. For example, one of the 22 age comparison condition variables is $x_1 - x_2 > -60$, 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 $x_1 - x_2 > -60$ is 1 because it is true that $35 - 32 > -60$. Another example is that the first person's age is less than or equal to 17: $x_1 \leq 17$, that is converted to the normalizing form of $-x_1 > -18$ in (1) with $a_1 = -1$, $a_2 = a_3 = 0$, and $b = -18$. Table 2 lists the 22 age comparison condition variables.

The following example illustrates how the age comparison variables are used to identify the edit rule of a householder's age being less than 15: $A_2^o = \{1\}$ (var2) and $A_{13}^o = \{1\}$ (var13). The normal form of the edit is

$$A_1 \times \{1\} \times A_3 \times \cdots \times A_{12} \times \{1\} \times A_{14} \times \cdots \times A_{31} = F$$

Another example is $A_5^o = \{3\}$ (var5), $A_{15}^o = \{1\}$ (var15), and $A_{16}^o = \{1\}$ (var16), in which the second person's hhr (var5) is child, the age (var15) is greater than 74, and the first person is less than 18 years older than the second person. In this example, the if-then-else edit rules in the 1997 ACS Edit and Allocation Specifications are

| | |
|---|---|
| Universe | *Child with Age* is greater than or equal to 75; |
| If... | *Age of Reference person* − *Age* is less than 18 and *Marital status* = Never married or SAS missing; |
| Then... | Blank *Age*; tally Z(12); set allocation flag. |

| | |
|---|---|
| Universe | *Child with Age* is greater than or equal to 75; |
| If... | *Age of Reference person* − *Age* is less than 18 and *Marital status* = Ever married; |
| Then... | Blank *Relationship*; tally Z(13); set allocation flag. |

The normal form of the edit is

$$A_1 \times \cdots \times A_4 \times \{3\} \times A_6 \times \cdots \times A_{14} \times \{1\} \times \{1\} \times A_{17} \times \cdots \times A_{31} = F$$

The age comparison also identifies 134 explicit edits, each of which is a contradiction condition within a subset of the 22 age comparison condition variables. For example, the normal form of the explicit edit

$$A_1 \times \cdots \times A_9 \times \{2\} \times \{1\} \times A_{12} \times A_{13} \times \{1\} \times A_{15} \times \cdots \times A_{31} = F$$

with $A^o_{10} = \{2\}$ (var10), $A^o_{11} = \{1\}$ (var11), and $A^o_{14} = \{1\}$ (var14) defines a contradiction situation among the variables var10, var11, and var14. If this edit is rewritten as the following inequalities:

$$\begin{array}{lrcrcrr} \text{var10:} & - & x_1 & & & \leq & -\ 18 \\ \text{var11:} & - & x_1 & + & x_2 & > & 14 \\ \text{var14:} & & & - & x_2 & > & -\ 15 \end{array}$$

it is clear that there are no values for $x_1$ (the first person's age) and $x_2$ (the second person's age) to satisfy the above three inequalities.

**Table 2. The 22 Age Comparison Condition Variables.**

| Variable ID | $a_1$ | $a_2$ | $a_3$ | $b$ | Variable ID | $a_1$ | $a_2$ | $a_3$ | $b$ |
|---|---|---|---|---|---|---|---|---|---|
| var10 | -1 | 0 | 0 | -18 | var21 | 0 | -1 | 1 | -4 |
| var11 | -1 | 1 | 0 | 14 | var22 | 1 | 0 | -1 | 14 |
| var12 | 1 | -1 | 0 | -60 | var23 | 0 | 1 | -1 | 14 |
| var13 | -1 | 0 | 0 | -15 | var24 | 0 | 1 | -1 | -15 |
| var14 | 0 | -1 | 0 | -15 | var25 | 1 | 0 | -1 | -15 |
| var15 | 0 | 1 | 0 | 74 | var26 | 0 | 0 | -1 | -30 |
| var16 | -1 | 1 | 0 | -18 | var27 | 0 | -1 | 0 | -30 |
| var17 | 0 | 0 | 1 | 74 | var28 | 0 | 0 | 1 | 59 |
| var18 | -1 | 0 | 1 | -18 | var29 | -1 | 0 | 1 | -35 |
| var19 | 1 | -1 | 0 | -1 | var30 | 0 | 1 | 0 | 59 |
| var20 | -1 | 0 | 1 | -4 | var31 | -1 | 1 | 0 | -35 |

## 4. The DISCRETE Edit Generation

The objective of the DISCRETE edit generation is to find a complete set of edits. A complete set of edits is the set of explicit (initially specified) edits and all essentially new implied edits derived from them. The main theorem of Fellegi and Holt demonstrated that, if one field in each failing edit (explicit or implicit) is changed, then the record with field values changed in the proper manner would pass all edits. The importance of a complete set of edits is illustrated with the example in Section 2, which has two explicit edits:

$$\begin{aligned} \boldsymbol{E}_1 = A^1_1 \times A^1_2 \times A^1_3 &= \{1\} \times \{2,3,4,5\} \times A_3 \\ \boldsymbol{E}_2 = A^2_1 \times A^2_2 \times A^2_3 &= A_1 \times \{1,3,4\} \times \{2\} \end{aligned}$$

Suppose we have a record $\boldsymbol{y} = (1, 2, 2)$, meaning a person, who is married and the spouse of the householder, has an age of less than 15, then $\boldsymbol{y}$ fails edit $\boldsymbol{E}_1$ and passes edit $\boldsymbol{E}_2$. In an attemp to correct the record, we list a single row matrix with entries 0 or 1, in which the entry is 1 if it is an entering field of

the failed edit, $\boldsymbol{E}_1$, and 0 otherwise:

$$\begin{array}{cc} \text{field} & \begin{array}{ccc} f_1 & f_2 & f_3 \end{array} \\ \boldsymbol{y} & \begin{pmatrix} 1 & 2 & 2 \end{pmatrix} \\ \boldsymbol{E}_1 & \begin{pmatrix} 1 & 1 & 0 \end{pmatrix} \end{array}$$

The sets of the field(s), $\{f_1\}$ and $\{f_2\}$, are two prime covers of the failed edit, $\boldsymbol{E}_1$. Therefore, we can change either $f_1$ or $f_2$, but not both, so that the new record is in $\overline{\boldsymbol{E}_1}$ and passes the edit, $\boldsymbol{E}_1$. If we decide to change $f_1$, we may choose a value from $\overline{A_1^1} = \{2\}$ and change the record to $\boldsymbol{y}_1 = (2, 2, 2)$. Alternatively, we may change $f_2$ and choose a value from $\overline{A_2^1} = \{1\}$ and change it to $\boldsymbol{y}_2 = (1, 1, 2)$. It is obvious that $\boldsymbol{y}_1 \in \overline{\boldsymbol{E}_1}$ and $\boldsymbol{y}_2 \in \overline{\boldsymbol{E}_1}$ and both of them pass $\boldsymbol{E}_1$. However, $\boldsymbol{y}_2 \in \boldsymbol{E}_2$, the only other explicit edit, and therefore fails $\boldsymbol{E}_2$ while $\boldsymbol{y}_1 \notin \boldsymbol{E}_2$ and passes $\boldsymbol{E}_2$.

To make sure the failing record, $\boldsymbol{y}$, being changed to $\boldsymbol{y}_1$ and not $\boldsymbol{y}_2$, we need additional information. This additional information comes from the so-called *implicit* or *implied edits*. 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 $\boldsymbol{E}_r$ are edits $\forall\ r \in S$, where $S$ is any index set,

$$\boldsymbol{E}_r : \prod_{j=1}^{n} A_j^r = F, \quad \forall \quad r \in S.$$

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

$$\boldsymbol{E}_* : \prod_{j=1}^{n} A_j^* = F \tag{2}$$

is an implied edit, where

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

$$A_i^* = \bigcup_{r \in S} A_i^r \neq \emptyset.$$

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 $\boldsymbol{E}_r$) $\forall\ r \in S$, but $A_i^* = A_i$, then the implied edit (2) 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 $\boldsymbol{E}_r$ $\forall\ r \in S$ from which the new implied edit $\boldsymbol{E}_*$ is derived are called *contributing edits*.

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

1. $A_j^* \neq \emptyset, \forall\ j,\ 1 \leq j \leq n$;
2. $A_i^r \neq A_i, \forall\ r \in S$, where $A_i^r \neq \emptyset$;
3. $A_i^* = A_i$.

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 (3).

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

$$\begin{aligned} &\text{Minimize} \quad \textstyle\sum_{r\in S} x_r \\ &\text{subject to} \quad \textstyle\sum_{r\in S} g_{rj}^i x_r \geq 1, \quad j = 1, 2, \cdots, n_i \qquad (3) \\ &\qquad x_r = \begin{cases} 1, & \text{if } \boldsymbol{E}_r \text{ is in the cover;} \\ 0, & \text{otherwise,} \end{cases} \\ &\qquad r \in S \end{aligned}$$

where

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

is the $j$th element in field $i$ of edit $\boldsymbol{E}_r$ ($r \in S$). If $\boldsymbol{x}$ is a prime cover solution to (3) 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 $i$th components cover all possible values of the entering field, which is the generating field to yield an essentially new implicit edit.

Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ are two cover solutions to (3) and $K_a = \{r \mid a_r = 1\}$ and $K_b = \{r \mid b_r = 1\}$. If $\boldsymbol{a}$ is a prime cover solution and $K_a$ is a proper subset of $K_b$, then the implied edit $\boldsymbol{E}_b$ derived from the contributing edits $\{\boldsymbol{E}_r \mid r \in K_b\}$ is redundant because $\boldsymbol{E}_b \subset \boldsymbol{E}_a$, which is derived from the contributing edits $\{\boldsymbol{E}_r \mid r \in K_a\}$. Therefore, prime cover solutions are more important and will generate nonredundant implicit edits. A redundant edit is an edit that is properly contained (as a subset) in another edit.

The simple example continues in this section. The field, $f_2$, is an entering field to both of $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ and $\{A_2^1, A_2^2\}$ is a prime cover of $A_2$. Furthermore, $A_1^1 \cap A_1^2 = \{1\} \cap A_1 = \{1\} \neq \emptyset$ and $A_3^1 \cap A_3^2 = A_3 \cap \{2\} = \{2\} \neq \emptyset$. Then, the edit

$$\boldsymbol{E}_3 = A_1^3 \times A_2^3 \times A_3^3 \ = \ \{1\} \times A_2 \times \{2\}$$

is an essentially new implicit edit. No other implicit edit can be derived from $\boldsymbol{E}_1$, $\boldsymbol{E}_1$, and $\boldsymbol{E}_3$. So the set, $\{\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3\}$, is a complete set of edits to our example. And record $\boldsymbol{y} = (1, 2, 2)$ also fails $\boldsymbol{E}_3$, the two-row matrix is listed as following:

$$\begin{array}{cc} \text{field} & \begin{array}{ccc} f_1 & f_2 & f_3 \end{array} \\ \boldsymbol{y} & \begin{pmatrix} 1 & 2 & 2 \end{pmatrix} \\ \boldsymbol{E}_1 & \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \\ \boldsymbol{E}_3 & \end{array}$$

The set, $\{f_1\}$, is the only prime cover solution to (4) in Section 6 of the failed edits $\boldsymbol{E}_1$ and $\boldsymbol{E}_3$. Therefore, we may change the value of $f_1$ to a value from $\overline{A_1^1 \cup A_1^3} = \{2\}$. The new imputed record $\boldsymbol{y}_1 = (2, 2, 2)$ passes all three edits. This formulation is called error localization and is the subject of Section 6.

## 5. Pre-Edits

The DISCRETE edit generation and the age comparisons are two major steps for the proposed methodology before the actual production is performed. The pre-edit step is the preparation for fitting the Fellegi-Holt model and performing the production when the data are available. The purpose of the pre-edits is to (1) identify the householder and spouse if present; (2) perform householder relationship pre-edits; (3) convert each of the households into a three-person household; (4) perform age and date of birth pre-edits and the consistency checks between age and date of birth; (5) perform race pre-edits; and (6) 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 householder relationship of the other persons in the same household has to be changed according to Table 3, 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".

We also 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. 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 inequalities defined in the age comparison condition variables in Section 3. With the available ACS data, there are 72375 individual records and 39407 at most three-person household records for the Fellegi-Holt modelling.

**Table 3. Householder 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 |

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 are

1. the birth month distribution is as following: 0.08385, 0.16183, 0.24520, 0.32633, 0.40912, 0.49026, 0.57576, 0.66284, 0.74868, 0.83517, 0.91600, 1.00000 (Wilcox [1999]);
2. age is unknown, birth year is known:
   (a) if known birth month, unknown birth day; generate the birth day from a conditional uniform distribution given the birth month and year;
   (b) if unknown birth month, known birth day; generate the birth month from a conditional distribution derived from Item 1 given the birth day and year;

(c) if unknown birth month; unknown birth day; generate the birth month from the distribution in Item 1 given the birth year; generate the birth day from a conditional uniform distribution given the generated birth month and the known birth year;

and compute the age from the date of birth and the response date;

3. age is known; birth year is known;

   (a) if known birth month, unknown birth day; generate the birth day from a conditional uniform distribution given the birth month and year and the response date;

   (b) if unknown birth month, known birth day; generate the birth month from a conditional distribution derived from Item 1 given the birth day and year and the response date;

   (c) if unknown birth month, unknown birth day; generate the birth month from the distribution in Item 1 given the birth year and the response date; generate the birth day from a conditional uniform distribution given the generated birth month and the known birth year and the response date;

   and compute the age from the date of birth and the response date; if the computed age and the reported age are not consistent, replace the reported age with the computed age;

4. age is known; birth year is unknown; compute the birth year from the age and the response date;

   (a) if known birth month, unknown birth day; generate the birth day from a conditional uniform distribution given the birth month and the computed year and the response date;

   (b) if unknown birth month, known birth day; generate the birth month from a conditional distribution derived from Item 1 given the birth day and computed birth year and the response date;

   (c) if unknown birth month, unknown birth day; generate the birth month from the distribution in Item 1 given the computed birth year and the response date; generate the birth day from a conditional uniform distribution given the generated birth month and the computed birth year and the response date;

   make adjustment of the birth year given the age, the reponse date, the birth month, and the birth day;

5. if both of the age and the birth year are missing, imputations of the age, the birth year, and possibly the birth month and the birth day are required;

6. if the reported age or the computed age is greater than 115, blank the age; an imputation of the age is required.

**Table 4. Donor Sequence for Race.**

| REL | REL definition | Donor Sequence |
|---|---|---|
| 1 | Householder | 5, 4, 3, 6, 2, 7, 8, 11, 10, 9, 12, 13 |
| 2 | Spouse | 7, 3, 1 |
| 3 | Child | 1, 3, 2 |
| 4 | Sibling | 1, 4, 5 |
| 5 | Parent | 1, 4, 5 |
| 6 | Grandchild | 3, 6, 1, 2 |
| 7 | In-Law | 2, 7, 1 |
| 8 | Other Relative | 8, 1, 2, 3, 4, 5, 6, 7 |
| 9 | Roomer or Boarder | 9, 13, 10, 11, 1 |
| 10 | Housemate or Roommate | 10, 1, 13, 11, 9 |
| 11 | Unmarried Partner | 3, 1, 13, 10, 9 |
| 12 | Foster Child | 12, 1, 2, 11, 13, 10, 9 |
| 13 | Other Nonrelative | 13, 1, 11, 2, 10, 9 |

The race pre-edits are to convert the race response to a three-digit race code. If the race response is missing and a donor in the household can be found, the donor's race will be used for imputation. Table 4 lists the donor sequence for race if a race response is missing. 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".

## 6. Error Localization

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 $\aleph = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \cdots, \boldsymbol{E}_m\}$ be a set of edits failed by a record $\boldsymbol{y}$ with $n$ fields, consider the set covering problem:

$$\begin{aligned} &\text{Minimize} && \textstyle\sum_{j=1}^{n} c_j x_j \\ &\text{subject to} && \textstyle\sum_{j=1}^{n} a_{ij} x_j \geq 1, \quad i = 1, 2, \cdots, m \end{aligned} \tag{4}$$

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

where

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

and $c_j$ is a measure of "confidence" in field $j$. We need to get $\aleph$ from a *complete* set of edits to obtain a meaningful solution to (4). A complete set of edits is the set of explicit edits (initially specified or replaced by a dominating implicit edit) and all essentially new implied edits derived from them.

If $\boldsymbol{x}$ is a prime cover solution to (4) and $K = \{r \mid x_r = 1\} \subset \{1, 2, \cdots, n\}$, then for each $k \in K$ we may change the value of $f_k$ to a value from

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

where $J = \{j \mid 1 \leq j \leq m,\ f_k \text{ is an entering field of } \boldsymbol{E}_j\}$. 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 an 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 $\aleph$. A simple example of the error localization was given in Section 4.

## 7. Discussion and Summary

We have discussed the edit methodology for the ACS study. The methodology is very effective if the given set of explicit sets is valid. A set of explicit edits, as specified by the subject matter experts, is called *invalid* if it is *inconsistent* and/or some important edits are missing. A set of edits is said to be *inconsistent* if they jointly imply that there are permissible values of a single field which would automatically cause edit failures. An *inconsistent* set of edits means that there is an implied edit $\boldsymbol{E}_r$ of the form

$$\boldsymbol{E}_r : A_1 \times \cdots \times A_{i-1} \times A_i^r \times A_{i+1} \times \cdots \times A_n = F,$$

where $A_i^r$ is a proper subset of $A_i$ for some Field $i$, i.e., $A_i^r \neq A_i$.

However, if $A_i^r$ is a subset of $A_i$, then this edit could not be an originally specified edit. Since the edit generating process identifies all implicit edits, it follows that this edit will also be generated. It is a simple matter to computer check the complete set of edits to identify implicit edits of this type and, thus, to determine whether the set of originally specified edits is inconsistent.

Missing edits may be critical when the imputation is performed. The set of explicit edits provides a fixed set of data points with the normal form. If a record is an element of the fixed set of data points, say $\boldsymbol{F}$, it fails some of the edits and is in need of some corrections. The corrected record must be an

element of $\overline{\boldsymbol{F}}$, the complement of the fixed set of data points. The corrected record might fail some of the missing edits, which are subsets of $\overline{\boldsymbol{F}}$.

In edit generation, the set covering problem (3) was formulated with integer programming, which is an optimization problem. In error localization, (4) was also a set covering problem (SCP). Both SCPs required a computationally efficient algorithm to obtain optimal solutions (or prime covers). Chen [1998] proposed a new set covering algorithm for the DISCRETE system. The new algorithm is at least 48 times faster with the two examples shown. It has been successfully implemented in the DISCRETE edit system.

It took 8 seconds for the pre-edit program to generate 39407 three-person household records from an input of 72375 individual records on a Sun Ultra machine. There were 31 fields in each of the three-person households. Nine of them were sex, marital status, and householder relationship; the other 22 were the age comparison condition variables. It took 53 seconds for the DISCRETE edit generation program to generate a complete set of 1154 edits on a Sun Ultra machine. The input included 31 fields and 163 explicit edits, in which 29 were identified from an ACS specification and 134 explicit edits were generated from the age comparison condition variables. It took 386 seconds for the error localization program to process the 39407 household records on a Sun Ultra machine. The age comparison program was run on a VAX/VMS machine and the running time was not measured.

## References


[1] B. Chen. Set covering algorithms in edit generation. In *Proceedings of the Statistical Computing Section*, pages 91–96. American Statistical Association, 1998.

[2] I. P. Fellegi and D. Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71:17–35, 1976.

[3] R. S. Garfinkel, A. S. Kunnathur, and G. E. Liepins. Optimal imputation of erroneous data: Categorical data, general edits. *Operations Research*, 34:744–751, 1986.

[4] A. J. Wilcox. Personal Communications, 1999.

[5] W. E. Winkler. Set-covering and editing discrete data. Technical report, Bureau of the Census, 1997.