USDA

United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Development
Division

RDD Research Report
Number RDD-02-06

October 2002

Elimination in Linear Editing and Error Localization

Stanley S. Weng

The views expressed herein are not necessarily those of NASS or USDA. This report was prepared for limited distribution to the research community outside the U.S. Department of Agriculture.

EXECUTIVE SUMMARY

To improve the computational efficiency in developing an editing and imputation (E/I) error localization module for the U.S. Census of Agriculture and other large surveys conducted by NASS, we addressed the methodological issue of variable elimination by equality edit in linear editing. This will simplify the linear edit system and reduce the magnitude of computation.

Though, in linear programming, the Fourier elimination method for linear inequalities has long been used, the role of equality edits in linear editing has not been fully explored. All the automatic computer E/I systems for numerical data have generally treated equality edits as a special case of inequality edits. A common practice has been to represent an equality edit by two inequalities of opposite direction. However, an equality edit defines a more informative relationship than an inequality edit. Therefore, the contribution of an equality edit to an editing problem should be more than that of an inequality edit.

Our research results, extending some of Fellegi and Holt (1976) results on linear edits, establish the methodology of variable elimination by equality edit in linear editing, which leads to a simplified linear editing problem in reduced dimension.

The methodological establishment of this paper can be particularly useful for the U.S. Census of Agriculture editing and imputation, for which a considerable number of the linear edits are equality ones. It is expected that the implementation of this methodology, in conjunction with other computational improvements, may enable Fellegi-Holt methodology to be implemented into the editing systems for future censuses and sample surveys with improved efficiency and accuracy.

RECOMMENDATIONS

The Census of Agriculture requires a very extensive editing system, featuring a large number of equality edits. As a result, the variable elimination methodology provided by this paper can be especially useful in the context of researching the possible incorporation of error-localization into the editing system for the 2007 Census of Agriculture. The following steps for further research are recommended:

1) Develop an automated approach for implementing the proposed variable elimination approach from an initial set of linear edits.
2) Calculate the computational gains from implementing this methodology on the linear edits prepared specifically for the Census of Agriculture.

ELIMINATION IN LINEAR EDITING AND ERROR LOCALIZATION

Stanley S. Weng

This paper presents some theoretical findings from our recent methodological research addressing the issue of variable elimination by equality edit in linear editing. The research was motivated by seeking improvement of computational efficiency for error localization, when implementing an error localization module for the editing and imputation of NASS' large surveys. Our results, extending some of Fellegi and Holt (1976) results on linear edits, establish the method of elimination by equality edit in linear editing, which leads to a simplified linear editing problem in reduced dimension.

The methodological establishment of this paper can be particularly useful as applied to the U.S. Census of Agriculture editing and imputation, for which a considerable number of the linear edits are equality ones. It is expected that the implementation of this methodology, in conjunction with other computational improvements, may enable Fellegi-Holt methodology to be implemented into the editing systems for future censuses and sample surveys with improved efficiency and accuracy.

KEY WORDS: Automatic editing and imputation; Fellegi-Holt methodology; Implied edit; Fourier elimination; Elimination by equality edit.

1. INTRODUCTION

For the error localization (EL) problem in automatic data editing and imputation (E/I) with linear edits under the Fellegi-Holt (F-H) methodology (Fellegi and Holt, 1976), the linear programming approach provides proper methods for solution (Rubin, 1975; Sande, 1978; Schiopu-Kratina and Kovar, 1989). However, in practice, the computational efficiency of error localization has been an issue (Winkler, 1999; Winkler and Chen, 2002). Various efforts have been made to improve the efficiency, including using an algorithm other than Chernikova's for linear programming, e.g., one based on Duffin's (1974) analysis of a system of linear inequalities (Houbiers, 1999); a tree-search approach instead of a Chernikova's algorithm-like process (Quere, 2000; Quere and De Waal, 2000); and even an entirely different approach, while still in the spirit of F-H (Bankier, 2000; Bankier, et al., 2000).

One other consideration is to simplify the linear edit system by using its special structure and features, to reduce the dimension of the system and thus the magnitude of computation for error localization.

Edits used in economic surveys and censuses, like those created by NASS/USDA for the U.S. Census of Agriculture, are primarily linear. They also contain a considerable number of equality edits, for example balance edits, in which an aggregate variable is equal to the sum of its component variables.

In the presence of equality edits in a linear edit system, it seems preferable to use the equality edits to eliminate fields (variables), leading to a simplified system in reduced dimension. However, until now, none of the

Stanley S. Weng is a Mathematical Statistician at the National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030. The author wishes to thank Dr. Ton de Waal, Statistics Netherlands, for his very helpful information and comments, provided in their correspondences, regarding the issues addressed in this paper. Thanks also to Dale Atkinson, Bob McEwen, Arnie Wilcox, and other reviewers at NASS for their suggestions which have helped to improve the presentation of this paper.

automatic computer E/I systems for numerical data have distinguished conceptually between equality and inequality edits. Equality edits have generally been treated as a special case of inequality edits. Some algorithms adopted the representation of an equality edit by two inequalities of opposite direction. Such handling seems to ignore the more informative specification of an equality edit. The equality form defines a more restrictive relationship than that of an inequality. In linear theory, an equality represents a lower dimension hyperplane in the data linear space. The contribution of an equality edit to an editing problem should be more than that of an inequality edit.

From the point of view of F-H methodology, there is an important distinction between equality and inequality edits in their generation of implied edits. This paper identifies such a distinction and establishes a method of using equality edits to eliminate fields and reach an equivalent linear edit system, for which all the inequality edits form a linear edit system of lower dimension. The original linear editing problem can be solved by first solving the problem with respect to this reduced system, and then determining the remaining fields by the specification of the equality edits.

Benefits in computational efficiency from this methodology can be significant. The magnitude of the editing problem is reduced through elimination, and the program needs only to handle inequality edits.

The outline of this report is as follows. Section 2 describes the basic setting and concepts of linear editing. Section 3 reviews some basic concepts and results of the F-H theory in the context of linear editing, that are related to the topic of this paper. Section 4 is a brief review of some mathematical concepts of Fourier elimination. Section 5 presents our theoretical results on the methodology of elimination by equality edit. Section 6 gets back to the main editing problem, error localization, which motivated this research and now can be solved in reduced scale with improved efficiency. Section 7 briefly discusses the implementation issue. Section 8 gives our recommendations. The technical Appendix contains the proof of the theoretical results of this paper.

2. LINEAR EDITING

The editing problem of numerical data from a survey/census is generally defined by a set of *linear edits* in the following form:

$$e_i: \ a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \le b_i \qquad i = 1,2,\ldots,m \qquad (1a)$$

with *positivity* constraints for the variables x_j:

$$x_j \ge 0, \ j = 1,2,\ldots,n \qquad (1b)$$

Here in (1a) the inequality sign may represent either inequality or equality. In matrix notation, the above linear edit system is written as

$$\mathbf{A}\,\mathbf{x} \le \mathbf{b} \qquad (2a)$$

and

$$\mathbf{x} \ge \mathbf{0} \qquad (2b)$$

where $\mathbf{A}$ $(m \times n)$ is the edit coefficient matrix of (1a), $\mathbf{b}$ $(m \times 1)$ is the right-hand-

side vector of (1a), and $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\tau$ is the data vector (where τ denotes transpose of a vector). Data editing so specified is called *linear editing*. Additional constraints may be added to the above basic setting to define various linear editing problems, for example error localization, that will be described in Section 6.

A data record is a *passing* record with respect to a linear edit system if the record satisfies all edits in the system. Otherwise, the record is a *failed* one. All data points that satisfy the linear edit system form the *feasible area* of the system. A passing record is also called feasible, and a failed record infeasible. A linear edit system is completely described by its feasible area. Two linear edit systems are considered equivalent if they have identical feasible areas. Geometrically, the feasible area of a linear system is a polyhedron in the data space.

We are actually in the setting of *linear programming* (Gass, 1985; Kotz & Johnson (Ed), 1985; Luenberger, 1984). Linear editing problems, such as error localization, are generally related to solutions of a linear program. A linear program can be solved by finding the set of all *extremal points* of its feasible area. Chernikova's algorithm (Chernikova, 1964, 1965) is used to find all extremal points of a linear system of nonnegative variables.

3. F-H THEOREM ON LINEAR EDITS

Fellegi and Holt (1976) established the fundamental theory of *automatic editing and imputation* in the following criteria, widely referred to as the F-H principles:

(1) The data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields).
(2) Imputation rules should be derived from the corresponding edit rules without explicit specification.
(3) When imputation takes place, it should maintain, as far as possible, the frequency structure of the data file.

For a failed record, identifying the fewest possible fields that may be changed to make the resulting record satisfy all edits is the *error localization* problem.

To solve the error localization problem, F-H showed that both explicit (the original) edits, as specified by subject-matter experts, and implied edits are needed. An *implied edit* is one that is logically implied by a set of explicit edits. An implied edit is said to be an *essentially new* edit if it does not involve all the fields (variables) explicitly involved in the edits that generated it. A field that is eliminated in generating an essentially new implied edit is called a *generating field* of the implied edit. A set of edits together with all essentially new implied edits that can be generated from the set of edits, forms a *complete set of edits*. The concept of a complete set of edits is crucial in F-H theory, which underlies their main theorem.

We focus on linear editing. For linear edits, the generation of essentially new edits and the derivation of a complete set of edits take an explicit form, as given by Theorem 3 of Fellegi and Holt (1976). The following is a restatement of the theorem.

Theorem (F-H, 1976). An essentially new implied edit e_t is generated from edits e_r and e_s, as in (1a), using field j as a generating field, if and only if a_{rj} and a_{sj} are both nonzero

and of opposite sign. The coefficients of the new edit, a_{tk}, are given by

$$a_{tk} = a_{sk}a_{rj} - a_{rk}a_{sj}, k = 1,2,\ldots,n,$$

where r and s are so chosen that $a_{rj} > 0$ and $a_{sj} < 0$. Repeated application of the above procedure will derive all essentially new implied edits.

The theorem simply states that from two linear inequalities where the inequality signs are in the same direction, a variable can be eliminated by taking their linear combination if and only if the variable has coefficients in the two inequalities which are of the opposite sign. The essence of generating an essentially new implied edit is elimination of a field.

Example 1 (Generation of essentially new implied edits). Consider the following set of linear edits:

E1: $2x + y \geq 20$,

E2: $x - 2y \geq 10$,

E3: $-3x + y \geq -60$,

E4 $-3x - y \geq -80$.

Using y as the generating field, the following essentially new edits may be generated:

$5x \geq 50$	(2*E1 + E2)
$-5x \geq -110$	(2*E3 + E2)
$-x \geq -60$	(E1 + E4)
$-6x \geq -140$	(E3 + E4)

The elimination operations are indicated in the parentheses. And, using x as the generating field, the following essentially new implied edits may be generated:

$5y \geq -60$	(3*E1 + 2*E3)
$-5y \geq -30$	(3*E2 + E3)
$y \geq -100$	(3*E1 + 2*E4)
$-7y \geq -50$	(3*E2 + E4).

We may continue to generate implied edits, though maybe redundant, from the above generated implied edits and the original edits.

4. FOURIER ELIMINATION

In linear theory, the method used in F-H Theorem 3 to generate essentially new implied edits is called *Fourier elimination* (Duffin, 1974; Fourier, 1826; Schrijver, 1986). This approach was proposed by Fourier to solve linear programming problems by elimination of variables. A variable, say, x_h, can be eliminated by taking positive combinations of two inequalities which have opposite signs in the coefficient of x_h. By adding suitable combinations of all possible pairs of inequalities with a positive and a negative coefficient of x_h, and subsequently adding all inequalities that did not contain x_h in the first place, one gets a new system of inequalities which does not contain variable x_h. This process can continue in successive elimination of other variables.

In a Fourier elimination process, the number of inequalities can grow excessively. Moreover, by taking all possible linear combinations of the original inequalities during the elimination process, it could easily occur that some inequalities become redundant. That is, an inequality can be written as a positive linear combination of some of the other inequalities. Duffin (1974), in his method of analyzing systems of linear

inequalities, proposed a "refined elimination" rule which deletes any inequality which has been generated by adding $t + 2$ or more of the original inequalities, when t variables have been eliminated. Houbiers (1999) applied Duffin's method to error localization.

Fourier's original problem of interest was whether a feasible solution to a specified set of linear inequalities exists. This can be restated, in the terminology of modern automatic data editing, as whether a set of fields can be imputed in such a way that a specified set of linear edits can be satisfied. Fourier's method of successive elimination has fostered modern automatic data editing, as generalized in the F-H methodology.

5. ELIMINATION BY EQUALITY EDIT

In addressing linear editing problems, it seems that the role of equality edits has not been fully explored. Equality edits have generally been treated as a special case of inequality edits, without using the defining feature, the deterministic aspect, of an equality edit. Actually, from the implied edit point of view, there is an important distinction between equality edits and inequality edits in their generation of implied edits, as shown by two lemmas to be introduced below.

Before stating the lemmas, we introduce the concept of equivalent edits. Two sets of edits are *equivalent*, if they imply each other, that is, each edit in one set is implied by (some edits of) the other set. In the linear edit context, two sets of linear edits are equivalent if their feasible area (thus, the set of extremal points) are identical. Two sets of equivalent linear edits have the same contribution to a linear edit system, and may thus replace each other. Editing problems with respect to two equivalent sets of edits are considered the same.

The following two lemmas extend the statements of Fellegi and Holt (1976) Theorem 3 in the situation where one edit is an equality. They state that, in such situations, it is always possible to generate an essentially new implied edit when a common field is involved. Furthermore, the original inequality edit can be replaced by the essentially new implied edit generated.

Lemma 1. An essentially new implied edit can always be generated from edits e_r and e_s, where e_s is an equality edit, using field j as a generating field, provided the coefficients of field j in the two edits are both nonzero.

Lemma 2. An inequality edit e_r can be replaced by an essentially new implied edit e_t generated from e_r and an equality edit e_s.

Proof of Lemma 1 and Lemma 2 are given in Appendix of this report.

The above lemmas show how an equality edit can be used to simplify a linear edit system. Based on these two lemmas, our next two theorems show that, just as elimination of free variables can be made using equalities in the linear system, so can elimination of positively constrained variables using the equality edits present in the linear edit system. The theorems are stated in the context of linear editing through the F-H concept of implied edit.

Theorem 1 (Elimination by equality edit). Suppose a linear edit system contains m inequality edits and one equality edit, with n positivity constraints for the n fields involved. Then, one nonzero field of the equality edit

can be eliminated from all other edits involving that field. The resulting new linear edit system contains $m+1$ inequality edits involving $n-1$ fields, with $n-1$ corresponding positivity constraints, and the original equality edit. The new system is equivalent to the original one. The extremal points of the original linear system can thus be obtained by first obtaining the extremal points in the $n-1$ fields of the new linear system excluding the equality edit, and then determining the remaining field by the equality edit.

Proof of Theorem 1 is given in Appendix. The following example illustrates the elimination method, as stated in the proof of Theorem 1.

Example 2. Consider the following set of linear edits:

$$\begin{aligned} 2x_1 + x_2 + x_3 &\le 4\,, \\ x_1 + 2x_2 + 3x_3 &\le 5\,, \\ x_1 + x_2 + 2x_3 &= 3\,, \\ x_j \ge 0,\ j &= 1,2,3\,. \end{aligned} \tag{4}$$

We use the equality edit to eliminate a variable, say, x_3, in the two inequality edits. Eliminating x_3 in the first inequality edit, we have

$$3x_1 + x_2 \le 5\,.$$

Eliminating x_3 in the second inequality edit, we have

$$-x_1 + x_2 \le 1\,.$$

And, the essentially new implied edit generated by the positivity constraint $x_3 \ge 0$ and the equality edit:

$$x_1 + x_2 \le 3\,.$$

In the (x_1, x_2) space, solve the last three inequalities, and there are three extremal points:

$$(5/3, 0)\,,\ (0,1)\,, \text{ and } (1,2)\,.$$

Now calculate x_3 using the equality edit, $x_3 = (3 - x_1 - x_2)/2$, it follows $x_3 = 2/3, 1, 0$, respectively. The extremal points of the original system thus are

$$(5/3, 0, 2/3)\,,\ (0,1,1)\,, \text{ and } (1,2,0)\,.$$

Theorem 1 may be extended to linear edit systems containing multiple equality edits, as follows.

Theorem 2. Suppose a linear edit system contains m inequality edits and q equality edits, with n positivity constraints for the n fields involved ($q \le n$). Assume the q equality edits are of full rank. Then, a new linear edit system, which is equivalent to the original one, can be formed through elimination using the q equality edits. The new system contains $m+q$ inequality edits involving $n-q$ fields, with $n-q$ corresponding positivity constraints, and the original q equality edits. The extremal points of the original linear system can thus be obtained by first obtaining the extremal points in the $n-q$ fields of the new linear system

excluding the q equality edits, and then determining the remaining q fields using the q equality edits.

Proof of Theorem 2 is provided in Appendix. The elimination process, as described in the proof of Theorem 2, is illustrated by the following example.

Example 3. Consider the following set of linear edits ($m = 1$ and $q = 2$):

$$3x_1 + 3x_2 + x_3 = 3\,, \qquad (5.1)$$
$$2x_1 + x_2 + 2x_3 = 4\,, \qquad (5.2)$$
$$4x_1 + 2x_2 + x_3 \le 3\,{}^1\!/_2\,, \qquad (5.3)$$
$$x_j \ge 0,\ j = 1,2,3\,.$$

Here, for a convenient setting to display the elimination process, we list equality edits above inequality edits.

First, use equality edit (5.1) to eliminate a field, say, x_3. With (5.2):

$$4x_1 + 5x_2 = 2\,. \qquad (5.4)$$

With (5.3):

$$x_1 - x_2 \le {}^1\!/_2\,. \qquad (5.5)$$

Also, with the positivity constraint $x_3 \ge 0$:

$$3x_1 + 3x_2 \le 3\text{, or}$$
$$x_1 + x_2 \le 1\,. \qquad (5.6)$$

Now the new edit system, resulting from the first stage of elimination, consists of two inequality edits, (5.5) and (5.6), involving fields (x_1, x_2), two corresponding positivity constraints $x_j \ge 0,\ j = 1,2$, and two equality edits, (5.1), and (5.4) (or, equivalently, the two original equality edits, (5.1) and (5.2)).

In the second stage of elimination, we use equality edit (5.4) to eliminate another field, say, x_2, in the inequality edits. With (5.5):

$$9x_1 \le {}^9\!/_2\text{, or,}$$
$$x_1 \le {}^1\!/_2\,. \qquad (5.7)$$

With (5.6):

$$x_1 \le 3 \qquad (5.8)$$

And with the positivity constraint $x_2 \ge 0$:

$$4x_1 \le 2\text{, or,}$$
$$x_1 \le {}^1\!/_2\,. \qquad (5.9)$$

The new edit system, resulting from the second stage of elimination, consists of three inequality edits, (5.7), (5.8) and (5.9), involving only field x_1, the positivity constraint $x_1 \ge 0$, and two equality edits (5.1) and (5.4) (or, (5.1) and (5.2)).

Now we can solve the reduced linear system in x_1, that is, (5,7), (5.8) and (5.9), with the constraint $x_1 \ge 0$. We find two extremal points $x_1 = 0$ and ${}^1\!/_2$. Then, using (5.4), we get $x_2 = {}^2\!/_5$ and 0 respectively; and using (5.1), $x_3 = {}^9\!/_5$ and ${}^3\!/_2$ respectively.

Thus, the extremal points of the original system are

$$(0, 2/5, 9/5) \text{ and } (1/2, 0, 3/2).$$

We may simultaneously eliminate two fields using the two equality edits. By this, we first convert the equalities to the canonical form through *Gaussian elimination* (see, e.g., Luenberger (1984)), and then substitute them into the inequalities. (5.1) and (5.2) can be written in such form in (x_1, x_2) as

$$x_1 + 5/3\, x_3 - 3 = 0\,, \qquad (5.10)$$

$$x_2 - 4/3\, x_3 + 2 = 0\,. \qquad (5.11)$$

Substituting (5.10) and (5.11) into the inequality edit (5.3) to eliminate x_1 and x_2, it follows

$$x_3 \geq 3/2\,; \qquad (5.12)$$

and into $x_1 \geq 0$, it follows

$$-5/3\, x_3 + 3 \geq 0\,; \qquad (5.13)$$

and into $x_2 \geq 0$, it follows

$$4/3\, x_3 - 2 \geq 0\,. \qquad (5.14)$$

Now the new edit system, resulting from the elimination of (x_1, x_2), contains three inequality edits in x_3, (5.12), (5.13) and (5.14), with the positivity constraint $x_3 \geq 0$, and the two equality edits, (5.10) and (5.11). We first solve the simplified linear system in x_3, obtaining the two extremal points, $x_3 = 3/2$ and $x_3 = 9/5$. Then, use (5.10) and (5.11) to determine the remaining fields, x_1 and x_2: for $x_3 = 3/2$, $x_1 = 1/2$ and $x_2 = 0$; and for $x_3 = 9/5$, $x_1 = 0$ and $x_2 = 2/5$.

6. ERROR LOCALIZATION

The error localization problem is stated as: for a failed record, anticipating the F-H principles, which components of the record must be changed in order that, with as few as possible changes, the record can be made to pass the edit system?

In linear editing, the linear programming approach to solving the error localization problem (Sante, 1978; Schiopu-Kratina and Kovar, 1989) is briefly described as follows.

Let $\mathbf{x}_0$ be a failed record with respect to the linear edit system (2a, 2b). Let $\Delta\mathbf{x}$ be the correction vector in the sense that $\mathbf{x}_0 + \Delta\mathbf{x}$ passes all the edits of (2a, 2b), that is,

$$\mathbf{A}(\mathbf{x}_0 + \Delta\mathbf{x}) \leq \mathbf{b}\,,$$
$$\mathbf{x}_0 + \Delta\mathbf{x} \geq \mathbf{0}\,.$$

Since $\mathbf{x}_0$ is known, rewrite the above system as

$$\mathbf{A}\Delta\mathbf{x} \leq \mathbf{b} - \mathbf{A}\mathbf{x}_0\,, \qquad (6)$$
$$\Delta\mathbf{x} \geq -\mathbf{x}_0\,.$$

The usual technique to solve (6) is to express the change $\Delta\mathbf{x}$ as a difference between

the positive and negative changes:

$$\Delta x = u - v \, ,$$

where both u and v are nonnegative vectors and their inner product is zero, $u^{\tau}v = 0$ (i.e., for any field, there may be either positive or negative change, but not both).

Denote $|x|^{+}$, called *cardinality*, for the number of strictly positive elements of a nonnegative vector x. The problem (6) can be stated as: Find all possible correction vectors $(u^{\tau}, v^{\tau})^{\tau}$ such that the cardinality of $|u - v|$ is a minimum, subject to:

$$\begin{aligned} A(u - v) &\le b - Ax_0 \, , \qquad (7) \\ u - v &\ge -x_0 \, , \\ u, v &\ge 0 \, , \\ u^{\tau}v &= 0 \, . \end{aligned}$$

Problem˙(7) can be restated with respect to a linear system in $(u, v) \in R^{2n}$, in standard form, with suitable matrices A_1 and b_1, as:

$$\min |u - v|^{+}$$

subject to:

$$A_1 \begin{pmatrix} u \\ v \end{pmatrix} \le b_1 \, , \qquad (8a)$$

$$u, v \ge 0 \, , \qquad (8b)$$

$$u^{\tau}v = 0 \, . \qquad (8c)$$

The complementary condition (8c) is actually redundant in the problem, because the minimum of $|u - v|^{+}$ is always reached at a point that satisfies (8c).

Rubin (1975) noticed the monotone property with Chernikova's algorithm in processing a row, that the cardinality of any new column generated is no less than that of its generating columns. He modified Chernikova's algorithm to solve the following *cardinality constrained linear program* problem:

$$\max d^{\tau}x$$

subject to

$$\begin{aligned} Ax &\le b \, , \qquad (9) \\ x &\ge 0 \, , \\ |x|^{+} &\le \eta \, , \end{aligned}$$

where x and d are $n \times 1$, A is $m \times n$, b is $m \times 1$, and η is a positive integer less than $\min\{m, n\}$, by directly producing the extremal points of the feasible area $G = \{x | Ax \le b, x \ge 0\}$ that satisfy $|x|^{+} \le \eta$, and then determining the optimal extremal point. As Tanahashi and Luenberger (1971) showed, an optimal solution to (9) can always be found in G .

Rubin's cardinality constrained linear program has been adopted as a standard formulation of the linear editing error localization problem, e.g., GEIS (Sande, 1978; Schiopu-Kratina and Kovar, 1989) and CherryPi (De Waal, 1996).

Houbiers (1999) applied Duffin's method on Fourier's analysis of linear inequality systems to error localization. He compared Duffin's method with Chernikova's algorithm - two similar algorithms with different control rules for excessive growth of the matrix, and

showed that Duffin's method is expected to be more efficient. Quere (2000) developed a new algorithm which performs Fourier elimination in a tree search process, instead of a Chernikova's algorithm-like process, to determine all optimal solutions to the error localization problem (see also Quere and De Waal, 2000).

In the presence of equality edits in the linear edit system, by the elimination methodology provided in last section, we can solve the error localization problem with respect to a simplified system in reduced dimension, as described below.

Through elimination by the equality edits, the linear edit system is restructured into the following form :

$$L_1: \quad \mathbf{A}_1\mathbf{x}^{(1)} \le \mathbf{b}^{(1)}, \quad \mathbf{x}^{(1)} \ge \mathbf{0},$$

and

$$L_2: \quad \mathbf{A}_2\mathbf{x} = \mathbf{b}^{(2)},$$

where $\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix} (n \times 1), \mathbf{x}^{(1)}\ (n-q)\times 1$ consisting of the fields involved in the inequality edits in $L_1, \mathbf{x}^{(2)}\,(q\times 1)$ consisting of the fields eliminated from the inequality edits; $\mathbf{A}_1\ m_1 \times (n-q), \mathbf{A}_2\,(q \times n)$ of full rank, $\mathbf{b}^{(1)}\,(m_1 \times 1)$, and $\mathbf{b}^{(2)}\,(q\times 1)$.

Let $\mathbf{x}_0 = \begin{pmatrix} \mathbf{x}_0^{(1)} \\ \mathbf{x}_0^{(2)} \end{pmatrix}$ be a failed record. The correction procedure is: if the subrecord $\mathbf{x}_0^{(1)}$ fails L_1, perform error localization and imputation for $\mathbf{x}_0^{(1)}$ with respect to system L_1. And then correct $\mathbf{x}_0^{(2)}$ by the imputed $\mathbf{x}_0^{(1)}$ using the equality edits of L_2. If $\mathbf{x}_0^{(1)}$ is feasible with respect to L_1, but $\mathbf{x}_0$ fails L_2, we only need to correct $\mathbf{x}_0^{(2)}$, again, by $\mathbf{x}_0^{(1)}$ using the equality edits of L_2, a deterministic imputation.

Benefits in computational efficiency for error localization can be significant from application of the elimination methodology. In processing a row with Chernikova's algorithm, excessive growth of the number of columns depends on the number of fields, which causes the storage problem. Reduction of the number of fields reduces the magnitude of computation. Also, the computer code does not need to handle equality edits, which also simplifies the computation.

7. IMPLEMENTATION

In linear editing, elimination of fields by equality edits restructures the linear edit system. This restructuring is conducted prior to data editing, since data are not involved. A separate module can be created to perform the elimination.

Generally, when q (linearly independent) equality edits are present in the linear edit system, any subset of q fields may be selected for elimination from the inequality edits, provided the elimination process is valid according to Theorems 1 and 2. That is, the q variables are linearly independent. When performing a successive elimination, at each stage, there is no additional theoretical criterion for choosing a field for elimination,

besides the general requirement of a nonzero field.

Practically, some strategies may be developed for choosing the fields for elimination. At each stage of elimination, maximizing the number of zeros in the coefficients of inequality edits appears to be a practical criterion. Aggregate variables are natural candidates for elimination. Other strategies may be developed based on the structure of the edit system.

In computer implementation of the elimination process, either successive elimination or simultaneous elimination can be performed, as illustrated by the examples in Section 5.

8. RECOMMENDATIONS

The edit specifications for many of NASS' surveys include a substantial number of equality edits. Balance edits are a common example of this type of edit. The Census of Agriculture, in particular, requires a very extensive editing system, incorporating many edits of this type. As a result, the variable elimination methodology provided by this paper is especially useful in the context of researching the possible incorporation of error-localization into the editing system for the 2007 U.S. Census of Agriculture.

The implementation of this methodology, in conjunction with other computational improvements, may enable Fellegi-Holt methodology to be implemented into the editing systems for future censuses and sample surveys with improved efficiency and accuracy.

The author makes the following recommendations for further research:

1) Develop an automated approach for implementing the proposed variable elimination process from an initial set of linear edits.

2) Calculate the computational gains from implementing this methodology on the linear edits prepared specifically for the Census of Agriculture.

REFERENCES

Bankier, M. (2000), "Imputing Numeric and Qualitative Variables Simultaneously," Technical Report, Social Survey Methods Division, Statistics Canada.

Bankier, M., Lachance, M., and Poirier, P. (2000), "2001 Canadian Census Minimum Change Donor Imputation Methodology - Extended Version of Report," Technical Report, Social Survey Methods Division, Statistics Canada.

Chernikova, N.V. (1964), "Algorithm for Finding a General Formula for the Non-negative Solutions of a System of Linear Equations," *USSR Computational Mathematics and Mathematical Physics*, **4**, 151-158.

Chernikova, N.V. (1965), "Algorithm for Finding a General Formula for the Non-negative Solutions of a System of Linear Inequalities," *USSR Computational Mathematics and Mathematical Physics*, **5**, 228-233.

De Waal, T. (1996), "CherryPi: A Computer Program for Automatic Edit and Imputation," Statistics Netherlands, Voorburg.

De Waal, T. (2000a), New Developments in Automatic Edit and Imputation at Statistics Netherlands. Report, Statistics Netherlands, Voorburg.

De Waal, T. (2000b), "An Optimality Proof of Statistics Netherlands' New Algorithm for Automatic Editing of Mixed Data," Report, BPA 3295-00-RSM, Statistics Netherlands, Voorburg.

De Waal, T.(2002), Personal correspondences.

Duffin, R.J. (1974), "On Fourier's Analysis of Linear Inequality Systems," *Mathematical Programming Studies*, Vol. I, 71-95, New York: North-Holland.

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

Fourier, J.B.J. (1826), "Solution d'une Question Particuliere du Calcul des Inegalites," Oeuvres II, Paris.

Houbiers, M.(1999), "Application of Duffin's Analysis of Linear Inequality Systems to the Error Localization Problem and Chernikova's Algorithm," Report, BPA 3107-99-RSM, Statistics Netherlands, Voorburg.

Gass, S.I. (1985), *Linear Programming*, Fifth Edition. New York: McGraw-Hill.

Kotz, S. and Johnson, N.L. (Edited) (1985), *Encyclopedia of Statistical Sciences*, Vol. 5, New York: John Wiley.

Luenberger, D.G. (1984), *Linear and Nonlinear Programming*, Second Edition, Reading, MA: Addison-Wesley.

Rubin, D.S. (1975), "Vertex Generation and Cardinality Constrained Linear Programs," *Operations Research*, **23**, 555-565.

Quere, R. (2000), "Automatic Editing of Numerical Data," Report, BPA 2284-00-RSM, Statistics Netherlands, Voorburg.

Quere, R., and De Waal, T. (2000), "Error Localization in Mixed Data Sets," Report, BPA 2285-00-RSM, Statistics Netherlands, Voorburg.

Sande, G. (1978), "An Algorithm for the Fields to Impute Problems of Numerical and Coded Data," Technical Report, Statistics Canada.

Sande, G. (1979), "Numerical Edit and Imputation," *Proceedings of the 42nd Session of the International Statistical Institute*, Manila, Philippines.

Schiopu-Kratina, I., and Kovar, J.G. (1989), "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Working Paper No. BSMD-89-001E, Statistics Canada. Ottawa, Ontario.

Schrijver, A. (1986), *Theory of Linear and Integer Programming*, New York: John Wiley.

Tanahashi, K. and Luenberger, D. (1971), Cardinality-Constrained Linear Programming, Stanford University.

Winkler, W.E. (1999), "State of Statistical Data Editing and Current Research Problems," Working Paper No. 29, Conference of European Statisticians. Rome, Italy.

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

APPENDIX: Proof of Convergence Theorems

Proof of Lemma [illegible]

[illegible]

APPENDIX: Proof of Lemmas and Theorems

Proof of Lemma 1:

The lemma is clearly true. Since we can always make the coefficient of the generating field in the equality edit to be opposite in sign to that in the other edit, the lemma is thus an immediate consequence of Fellegi and Holt (1976) Theorem 3.

Proof of Lemma 2:

The set of edits e_r and e_s is equivalent to the set of edits e_t and e_s, since edit e_r can also be generated as an implied edit by edits e_t and e_s. Thus we may use the set of edits e_t and e_s to replace the original set of edits e_r and e_s; or, equivalently, use the essentially new implied edit e_t to replace the original inequality edit e_r.

Proof of Theorem 1:

Let $e_i, i = 1,2,\ldots, m$, be the m inequality edits of the linear edit system, and $\overline{e}$, the one equality edit. Denote g_j for the positivity constraint $x_j \geq 0, j = 1,2,\ldots, n$.

Suppose x_h is a nonzero field of $\overline{e}$. For each inequality edit e_i, for which x_h is also a nonzero field, by Lemmas 1 and 2, an essentially new implied edit can be generated from e_i and $\overline{e}$ using field x_h as the generating field, and replaces e_i in the original set of edits. Also, an essentially new implied edit can be generated from g_h, the positivity constraint for x_h, and $\overline{e}$ using x_h as the generating field, and replaces g_h. The resulting linear edit system is equivalent to the original system. The new system contains $m+1$ inequality edits in which x_h is eliminated, with $n-1$ positivity constraints $g_j, j \neq h$, and the original equality edit $\overline{e}$. In the new linear system, $\overline{e}$ is the only edit that involves x_h. Formally, x_h is a free variable in $\overline{e}$. Thus, the extremal points of the new linear system can be obtained by first obtaining the extremal points in $x_j, j \neq h$, of the new linear system excluding $\overline{e}$, and then determining x_h using $\overline{e}$. These extremal points are also those of the original linear system. The proof is completed.

Proof of Theorem 2:

This theorem is a result of repeated application of the elimination method given by Theorem 1. For convenience, the proof is made for $q = 2$. For $q > 2$, the proof can be formally given by induction, and is omitted here.

Denote L for the linear edit system in the theorem. Let $e_i, i = 1,2,\ldots, m$, be the m inequality edits, $\overline{e}_k, k = 1,2$, the two equality edits. Denote g_j for the positivity constraint $x_j \geq 0, j = 1,2,\ldots, n$.

Suppose x_h is a nonzero field of the equality edit $\overline{e}_1$. By Theorem 1, using $\overline{e}_1$ we can eliminate field x_h from all other edits that involves x_h, including the m inequality edits, $e_i, i = 1,2,\ldots, m$, the equality edit $\overline{e}_2$, and

the positivity constraint g_h. Denote $e_i^{(1)}$ for the resulting m inequality edits from $e_i, i = 1,2,\ldots, m$, $\overline{e}_2^{(1)}$ for that from $\overline{e}_2$, and $e_{m+1}^{(1)}$ for that from g_h. Denote $L^{(1)}$ for the resulting linear edit system. $L^{(1)}$ then contains $m+1$ inequality edits, $e_i^{(1)}$, $i = 1,2,\ldots, m+1$, which involve fields $x_j, j \neq h$, the $n-1$ corresponding positivity constraints, $g_j, j \neq h$, and the two equality edits, $\overline{e}_1$ and $\overline{e}_2^{(1)}$ (one original and one generated). $L^{(1)}$ is equivalent to L. This is the first stage of elimination.

Now we perform the second stage of elimination with respect to $L^{(1)}$. Let $x_{h'}$ be a nonzero field of the equality edit $\overline{e}_2^{(1)}$ ($h' \neq h$, h' exists by the full rank assumption for the equality edits). We use $\overline{e}_2^{(1)}$ to eliminate field $x_{h'}$ from all other edits, except $\overline{e}_1$, that involves $x_{h'}$, including the $m+1$ inequality edits, $e_i^{(1)}$, $i = 1,2,\ldots, m+1$, and the positivity constraint $g_{h'}$ for $x_{h'}$. Denote $e_i^{(2)}$, $i = 1,2,\ldots, m+2$, for the resulting $m+2$ inequality edits.

Denote $L^{(2)}$ for the resulting linear edit system from the second stage of elimination. $L^{(2)}$ contains $m+2$ inequality edits, $e_i^{(2)}, i = 1,2,\ldots, m+2$, involving fields $x_j, j \neq h, h'$, the $n-2$ corresponding positivity constraints, $g_j, j \neq h, h'$, and the two equality edits, $\overline{e}_1$ and $\overline{e}_2^{(1)}$ (or equivalently, the original set of equality edits $\overline{e}_1$ and $\overline{e}_2$). $L^{(2)}$ is equivalent to $L^{(1)}$, and hence to L.

We thus established, for $q = 2$, the structure of the new linear system through elimination, as stated in the theorem. The general truth of the theorem for any q can be established by induction. The statement in the theorem for obtaining the extremal points of the original system through the new system is an immediate consequence of the structure of the new linear system. The proof is completed.