Reduction of Linear Edits*

Stanley Weng, NASS/USDA
3251 Old Lee Hwy, Room 305, Fairfax, VA 22030

KEY WORDS: Linear Editing, Fellegi-Holt Methodology, Polyhedral Theory, Implicit Equality, Elimination by Equality, Facet Approach

1. Introduction

1.1. Linear editing

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

$$a^i x \le b_i, i = 1,2,\ldots,m \qquad (1.1a)$$

where $a^i = (a^i_1,\ldots,a^i_n)$ is the coefficient vector of the ith edit, and $x = (x_1,\ldots,x_n)^T$ (T denotes transpose of a matrix) is the data record vector with *positivity* constraints for variables x_j:

$$x_j \ge 0,\ j = 1,\ldots,n \qquad (1.1b)$$

In (1.1a) the inequality sign may represent either inequality or equality. In matrix notation, the above linear edit system is written as

$$\begin{aligned} Ax &\le b \\ x &\ge 0 \end{aligned} \qquad (1.2)$$

where $A,\ m \times n$, having rows $a^i, i = 1,\ldots,m$, is the coefficient matrix of (1.1a), and $b = (b_1,\ldots,b_m)^T$ is the right-hand-side vector of (1.1a). Data editing so specified is called *linear editing*. Additional constraints may be added to the above basic setting to define various linear editing problems. 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. A passing record is also called *feasible*, and a failed record *infeasible*. All data points $x \in R^n$ that satisfies a linear edit system form the *feasible area* of the system. Denote it as S, thus

$$S = \{x \in R^n : Ax \le b, x \ge 0\}. \qquad (1.3)$$

As by convention, we do not distinguish the linear edit system (1.2) and its feasible area S (1.3), and use S to denote both the edit system and its feasible area. A linear edit system is completely described by its feasible area. Two linear edit systems are considered equivalent if they have the same feasible area.

We are actually in the setting of *linear programming* (Gass, 1985; Luenberger, 1984; Nemhauser and Wolsey, 1988; Schrijver, 1986). Linear editing problems are generally related to solution of a linear program. Linear programming techniques have been used to address linear editing problems (Giles, 1989; Houbiers, 1999; Quere, 2000; Quere and De Waal, 2000; Rubin, 1975; Schiopu-Kratina and Kovar, 1989; Tanahashi and Luenberger, 1971).

1.2. Fellegi-Holt theory on linear edits

Fellegi and Holt (referred as F-H) (1976) established the fundamental theory of automatic editing and imputation. They introduced the following basic concepts for the theory, leading to their profound theoretical results (Theorem 1 and the corollaries) of the F-H methodology. For a insightful review of the F-H methodology, see Winkler (1999) and Winkler and Chen (2002).

An edit that is logically implied by some other edits, is called an *implied edit* of the other edits. Those other edits are called the *generating edits* of the implied edit. An implied edit is said to be an *essentially new implied edit* if it does not involve all the fields 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 them forms a *complete set of edits* (with respect to the essentially new implied edits). Two sets of edits are called *equivalent*, if they imply each other, that is, each edit in one set is implied by some edits of the other set (for linear edits, just have the same feasible area). Editing problems with respect to two equivalent sets of edits are considered the same.

For linear edits, positive linear combinations of linear edits generate implied edits. In the context of linear editing, the generation of essentially new edits and derivation of a complete set of edits take explicit form, as given by F-H (1976), Theorem 3.

1.3. Reduction of linear edits

In a linear edit system S, the linear edits merely provide a description of S. The set of linear edits originally used to define S may not be the simplest in the sense that the system may be replaced by another equivalent system containing fewer edits or involving fewer variables. Any specification for S that is logically implied by other

specifications in S is considered not necessary, or *redundant*, and can be removed from the system without changing the feasible area.

In application, to enter an automatic linear editing system, the set of linear edits needs to be "minimal" with all redundances removed. In linear editing, edit analysis is conducted to identify redundances. Giles (1989) described the types of edit analysis to determine the minimal set of edits required to define the feasible area of the system, and intuitively justified the methods used for edit analysis. However, the analysis does not provide insight into the structure of the linear system, nor further characterization of the redundance besides the descriptive definition.

Characterization and identification of redundance in different forms with a linear edit system is the issue of reduction of linear edits. This paper develops reduction techniques of linear edits. A systematic approach, called the *facet approach*, is developed to reach the minimal representation of a linear edit system, based on the representation theory of linear inequalities and using some techniques developed in this paper.

Section 2 of this paper reviews the representation theory of linear inequalities. Section 3 discusses reduction of a linear edit system by its equality set. The concept of implicit equality is introduced, and linear-equation type characterization is developed for implicit equalities. Section 4 discusses reduction of the linear system in its inequality set. The facet approach is proposed. For the purpose of this paper, proofs of the theoretical results are not contained here.

2. Minimal representation of linear edits

To clarify the notion of necessary description for the feasible area of a linear edit system, we need some mathematical knowledge of linear inequalities - the *polyhedral theory* (Nemhauser and Wolsey, 1988). A *polyhedron* $P \subseteq R^n$ is the set of points that satisfy a finite number of linear inequalities,

$$P = \{x \in R^n : Ax \le b\} \qquad (2.1)$$

where A, $m \times n$, has rows $a^i, i = 1,\ldots,m$, and $b = (b_1,\ldots,b_m)^T$. In this context, the matrix (A,b) is considered as a representation of P. A member inequality $a^i x \le b_i$ of P might appear as an inequality but actually be satisfied at equality by all points of P. We conceptually classify the two groups of member inequalities of P. The rows (a^i, b_i) of (A,b), satisfying $a^i x = b_i$ for all $x \in P$, consist in the *equality set* of P, denoted as $(A^=, b^=)$. And the other rows of (A,b) consist in the *inequality set* of P, denoted as $(A^\le, b^\le)$. This conceptual classification of equality and inequality sets is important in the development of representation theory of polyhedra.

A member inequality $a^i x \le b_i$ of P represents a *face* of P, $F = \{x \in P : a^i x = b_i\}$. Each face F of P is itself a polyhedron. A special kind of faces, called *facet*, is of particular importance for the description of P. A face F of P is called a facet if $\dim(F) = \dim(P) - 1$, where $\dim(P)$ denotes the dimension of $P \subseteq R^n$, which is determined by $\dim(P) = n - \text{rank}(A^=, b^=)$, where $\text{rank}(\,)$ denotes the rank of a matrix. If $\dim(P) = n$, P is of full-dimension. If F is a facet of P, there exists some inequality $a^r x \le b_r$ in the inequality set $(A^\le, b^\le)$ representing F. For each facet F of P, one member of the equivalence class of inequalities representing F is necessary for the description of P. The facets are also sufficient for the description of P. An inequality $a^r x \le b_r$ in the inequality set $(A^\le, b^\le)$ that represents a face of P of dimension less than $\dim(P) - 1$ is irrelevant to the description of P.

The minimal representation theorem of polyhedra states that a full-dimensional polyhedron has a unique (to within scalar multiplication) minimal representation by a finite set of linear inequalities, each of which represents a facet. When the equality set is not empty, P is not full-dimensional, and the minimal representation of P consists of a maximal linearly independent subset of the equality set, and a set of inequalities each of which represents a facet. And, in this situation, a linear combination of the equality edits may be attached to the members in each equivalence class of inequalities of facets.

We have a clear picture for a polyhedron P. Facets are essential in the description of P. The equality set

of P, if not empty, constrains P to a lower dimension subspace and plays no essential role for the description of P in the sense that it only contributes some redundant members to each equivalence class of inequalities of facets. The equality set thus represents the redundance in dimension. With $\text{rank}(A^{=}, b^{=}) > 0$, the polyhedron P in R^n is not full-dimensional and can be equivalently presented in a h-dimensional subspace, where $h = \dim(P) = n - \text{rank}(A^{=}, b^{=})$, as a full-dimensional polyhedron.

For a polyhedron P, reduction related to the equality set is to reduce the dimension of the space where P resides, to make P full-dimensional. Reduction related to the inequality set is to remove irrelevant inequalities, that is, those representing a face of dimension lower than $\dim(P) - 1$, and keeps only one member for each equivalence class of inequalities representing a facet. In the following we discuss the two aspects of reduction respectively.

3. Reduction of edits by the equality set

For equality edits in the system, elimination by equality (Weng, 2002) can be performed to reduce the linear edit system to an equivalent system for which all the inequality edits form a linear edit system of lower dimension.

3.1. Implicit equality

The equality set has been, however, defined descriptively, without indicating how to identify the equality set for a given edit system S. In application of obtaining minimal representation of a linear edit system, it is an issue to identify the equality set. A member inequality might be actually satisfied at equality by all points in S. We call such a member *implicit equality* with respect to the set of edits. The equalities originally presented in the system may be called explicit equality.

In the following we give characterizations of implicit equalities. We assume the system S, as in (1.3), does not contain explicit equalities - their presence in the system will hide the implicit equalities. Suppose S is not empty.

The following proposition, though simple, introduces the characterization of implicit equalities through a nonnegative (row) m-vector λ.

PROPOSITION 1. *The system S contains implicit equalities if and only if there is a nonnegative vector $\lambda \in R_+^m$, $\lambda \neq 0$, such that for $x \in S$*

$$\lambda(Ax - b) = 0. \qquad (3.1)$$

We are looking for general characterization of implicit equalities involving only the matrix representation (A, b) of S. The following is one based on the duality of linear programming.

THEOREM 1. *The r-th inequality $a^r x \leq b_r$ of the system S is an implicit equality if and only if there is a nonnegative vector $\lambda \in R^m$, $\lambda \geq 0$ such that*

$$\lambda A \geq 0 \qquad (3.2)$$
$$\lambda b = 0,$$

and the r-th component of λ is strictly positive.

If we only need to confirm the existence of implicit equalities in the system, but not necessary to identify which ones, we need to show that the feasible area of the following linear system in λ is not empty (except the null point 0):

$$\lambda A \geq 0 \qquad (3.3)$$
$$\lambda b = 0$$
$$\lambda \geq 0.$$

With the additional condition that none of the variables x_i is always zero on S, which is generally satisfied in practice, Theorem 1 can be enhanced to that the first inequality of (3.2) becomes an equation, that is, a linear equation characterization for implicit equalities, as the next theorem states.

THEOREM 2. *Suppose for the system S none of the variables $x_j = 0$ for all $x \in S$. Then the i-th inequality $a^i x - b_i \leq 0$ of S is an implicit equality if and only if there is $\lambda \in R_+^m$ with $\lambda_i > 0$ satisfying*

$$\lambda A = 0 \qquad (3.4)$$
$$\lambda b = 0.$$

For any nonnegative vector $\lambda \in R_+^m$ satisfies (3.4), the set $I_\lambda = \{i: \lambda_i > 0\}$ identifies implicit

equalities $a^i x \le b_i$, $i \in I_\lambda$. If we only need to confirm the existence of implicit equalities in the system, but not necessary to identify which ones, we need to show that the feasible area of the following linear system in $\lambda \in R^m$ is not empty (except the null point 0):

$$\lambda(A,b) = 0, \lambda \ge 0.$$

In some situations we may limit the searching for implicit equalities to a subset of the member inequalities in the system, or conclude that there are no implicit equalities in the system by simple criterion, as the following corollary states.

COROLLARY 1. *Under the condition of Theorem 2 of implicit equality, if in the system* S, b *is nonnegative, let* $I_b = \{i: b_i = 0\}$, *then only for* $i \in I_b$, *the inequality* $a^i x \le b_i$ *is a candidate of implicit equality. In particular, if* $I_b = \varnothing$ (i.e., $b > 0$), *then the system (3.3) does not contain any implicit equality. If* $I_b \ne \varnothing$ *and there is a column* $a_{(j)}$ *of* A, *for which the i-th element* $a^i_{(j)} > 0$ *for all* $i \in I_b$, *then the system does not contain any implicit equality.*

Corollary 1 can be extended to a subset of columns of (A, b), if they have such a "covering property" as stated in the next corollary.

COROLLARY 2. *Under the condition of Theorem 2 of implicit equality, if there is a subset of the columns of* (A, b), *say* $\{\tilde{a}_{(l)}, l = 1, \ldots, t\}$, *having such property that* $\tilde{a}_{(1)}$ *is nonnegative, and let* $I_1 = \{i: \tilde{a}^i_{(1)} = 0\}$; $\tilde{a}_{(2)}$ *is nonnegative on* I_1, *and let* $I_2 = \{i \in I_1: \tilde{a}^i_{(2)} = 0\}$ *; and so on, finally,* $\tilde{a}_{(t)}$ *is nonnegative on* $I_{t-1} = \{i \in I_{t-2}: \tilde{a}^i_{(t-1)} = 0\}$, *and let* $I_t = \{i \in I_{t-1}: \tilde{a}^i_{(t)} = 0\}$. *Then only for* $i \in I_t$, *the inequality* $a^i x \le b_i$ *is a candidate of implicit equality. If* $I_t = \varnothing$, *the system contains no implicit equalities.*

Corollary 2 opens the chance to considerably reduce the candidate set of implicit equalities. If I_t is identified, so $\lambda_i = 0$ for $i \notin I_t$, then the equation $\lambda(A,b) = 0$, or $\sum_{i=1}^{m} \lambda_i (a^i, b_i) = 0$, reduces to $\sum_{i \in I_t} \lambda_i (a^i, b_i) = 0$. In the set of linear edits for the U.S. Census of Agriculture, the constants b_i are usually nonnegative.

3.2. Projective Algorithm for Feasibility Problem

If, after possible reduction, the candidate set of implicit equalities can not be concluded empty, we need to solve such a linear program

$$\lambda(A.b)_c = 0$$
$$\lambda \ge 0$$

where $(A,b)_c$ is the candidate set of implicit equalities. Such a linear programming problem is called the *homogeneous feasibility problem*, generally stated as: Let $Q^0 = \{r \in R^n: Gr = 0, r \ge 0\}$, where G, $m \times n$, with $\mathrm{rank}(G) = m$. Find a ray $r \in Q^0 \setminus \{0\}$ or show that $Q^0 = \{0\}$ (empty). There is a *projective algorithm* available to solve the linear program (see Nemhauser and Wolsey, 1988, I.6.4).

4. Reduction of edits in the inequality set

Suppose the linear system $S: Ax \le b, x \ge 0$, consists of only the inequality set, that is, $(A^=, b^=) = \varnothing$, as after elimination of all (explicit and implicit) equality edits. S is thus of full rank.

4.1. Implied edit

First give a general characterization for implied edits in a linear edit system based on the duality of linear programming.

PROPOSITION 2. *If the linear edit system* S *is not empty, the edit* $a^r x \le b_r$ *is an implied edit for the system if and only if there is* $\lambda \in R^m$ *with*

$$\lambda_r = -1 \text{ and } \lambda_i \ge 0 \text{ for } i \ne r \quad (4.1)$$

such that

$$\lambda A \ge 0 \quad (4.2)$$

and

$$\lambda b \le 0. \quad (4.3)$$

Positive linear combinations of linear edits generate implied edits. Such a notion has often been perceived as an alternative to the descriptive definition of implied edits. The following corollary gives an accurate statement.

COROLLARY 3. *If the system* S *is not empty, then the edit* $a^r x \le b_r$ *is an implied edit for the system if and only if*

$$a^r x - b_r \le \sum_{i \ne r} \lambda_i (a^i x - b_i),$$
$$\lambda_i \ge 0, i \ne r. \quad (4.4)$$

(4.4) shows the implied edit $a^r x \le b_r$ is dominated by a linear positive combination of (or, *linearly dominated* by) other inequality edits in the system.

4.2. The facet approach

Suppose the linear edit system $S = \{x \in R^n : Ax \le b, x \ge 0\}$ has its equality set empty. By the polyhedral theory, the inequalities are classified into two categories: one for those that represents a facet and the other for those that does not, according to the dimension of the face the inequality represents being $\dim(S) - 1$ or less.

We need to identify one representative inequality for each facet and remove all other inequalities. We propose an approach as follows. For each inequality $a^i x \le b_i$ in S, if the face it represents, $F_i = \{x \in S : a^i x = b_i\}$, is a facet, then $\dim(F_i) = \dim(S) - 1$, so $\text{rank}(A^{=}_{F_i}, b^{=}_{F_i}) = 1$, that is, the equality set of face F_i contains only one member, the defining equality of the face. Otherwise, if the equality set of F_i contains other linearly independent members than the defining equality, then $\dim(F_i) < \dim(P) - 1$, F_i is not a facet. Here, again, identification of implicit equalities is involved. We proceed in such way: first performing elimination by the equality $a^i x = b_i$ for face F_i and then examine if the face contains any implicit equalities. If not, then F_i is a facet. We then further remove other inequalities in S, if any, representing the same facet. If F_i is not a facet, the inequality $a^i x \le b_i$ together with all its multiplications is to be removed from the system We call such an approach the *facet approach*.

The facet approach directly classifies the faces for facet based on the polyhedral representation theory, avoiding unnecessary involvement of other irrelevant structures, for example, linear combinations of member inequalities.

Acknowledgments

The author wishes to thank his colleagues at NASS/ USDA for their suggestions which have helped to improve the presentation of this paper.

References

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.

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

Giles, P. (1989). "Analysis of Edits in a Generalized Edit and Imputation System," Working Paper No. SSMD-89-004 E, Statistics Canada, Ottawa, Ontario.

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.

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.

Nemhauser, G.L., and Wolsey, L.A.(1988), *"Integer and*

Combinatorial Optimization, John Wiley, New York.

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.

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.

Weng, S.S. (2002), "Elimination in Linear Editing and Error Localization," *2002 Proceedings of the American Statistical Association, Survey Research Methods Section* [CD-ROM], 3685-3690, Alexandria, VA: American Statistical Association.

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.

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