Dr. 2190

216
1/14/81
T.S.
R1170

(2)

ornl

ORNL/TM-7126/R1

MASTER

**OAK
RIDGE
NATIONAL
LABORATORY**

**UNION
CARBIDE**

A Rigorous, Systematic Approach to Automatic Data Editing and Its Statistical Basis

G. E. Liepins

**OPERATED BY
UNION CARBIDE CORPORATION
FOR THE UNITED STATES
DEPARTMENT OF ENERGY**

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A03; Microfiche A01

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Contract No. W-7405-eng-26

A RIGOROUS, SYSTEMATIC APPROACH TO AUTOMATIC DATA EDITING
AND ITS STATISTICAL BASIS

G. E. Liepins

Regional and Urban Studies Section

Energy Division

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or approval by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Date Published - January 1981

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37830
operated by
UNION CARBIDE CORPORATION
for the
DEPARTMENT OF ENERGY

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

CONTENTS

ABSTRACT	1
1. INTRODUCTION	3
2. CONNECTIONS TO MAXIMUM POSTERIOR PROBABILITY	4
3. THEORETICAL DETAILS OF THE FELLEGI-HOLT APPROACH	9
4. THE SET-COVERING PROBLEM	16
5. MODIFICATIONS OF THE WEIGHTING	19
6. DETAILS OF INVARIANCE	22
7. COMPUTATIONAL DETAILS FOR DISJOINT SUFFICIENT COLLECTIONS	24
8. CONCLUSIONS	26
REFERENCES	28

*A RIGOROUS, SYSTEMATIC APPROACH TO AUTOMATIC DATA EDITING
AND ITS STATISTICAL BASIS*

G. E. Liepins

ABSTRACT

Automatic data editing is the computerized identification and correction (optional) of data errors. These techniques can provide error statistics that indicate the frequency of various types of data errors, diagnostic information that aids in identifying inadequacies in the data collection system, and a "clean" data base appropriate for use in further decision making, in modeling, and for inferential purposes. However, before these numerous benefits can be fully realized, certain research problems need to be resolved, and the linkage between statistical error analysis and extreme-value programming needs to be carefully determined.

The linkage is provided here for the special case that certain independence and symmetry conditions obtain; also provided are rigorous proofs of results central to the functioning of the Boolean approach to automatic data editing of coded (categorical) data. In particular, sufficient collections of edits are defined, and it is shown that for a fixed objective function the solution to the fields to impute problem is obtainable simply from knowing which edits of the sufficient collection are failed, and this solution is invariant of the particular sufficient collection of edits identified. Similarly, disjoint-sufficient collections of edits are defined, and it is shown that, if the objective function of the fields to impute problem is determined by what Freund and Hartley call the "number of involvements in unsatisfied consistency checks," then the objective function will be independent of the disjoint-sufficient collection of edits used.

**THIS PAGE
WAS INTENTIONALLY
LEFT BLANK**

1. INTRODUCTION

The identification and correction of data errors is of interest to both theoreticians and data users, and one of the tools recognized to be useful in such investigations is automatic data editing, the computerized location and correction (optional) of data errors. With the widespread advent of the computer, automatic data editing naturally became a focal point of interest; presumably, early hopes were that the availability of powerful computing capabilities would make the screening of massive data sets a routine matter. Not unexpectedly, some of the early applications began on an ad hoc basis, without a sound theoretical foundation, and it was only recently that the problem was attacked of how to systematically and efficiently infer which fields (coordinates) of an erroneous multivariate record were responsible for the record's failure and therefore needed to be changed. Two papers contributing to the underpinnings of the theory are those by Naus, Johnson, and Montalvo¹ and Fellegi and Holt.² Although the Fellegi-Holt work does not specifically state the conclusion, an important property of their method is the independence of the eventual solution with regard to the specific representation of the edits (constraint conditions). Some of the previously suggested methods do not have this vital property.

Recently, additional theory has been developed — especially by Sande³ and McKeown.⁴ Most of their work has been directed at solving the fields to impute problem for continuous data subject to linear constraints, and the results have been presented in the vernacular of extreme-value programming; little investigation has been made into the statistical determination of the appropriate objective function of the

resultant extreme-value problem or into the comparison of these techniques with those of Fellegi and Holt. Liepins⁵ suggested that, with appropriate modifications, the Fellegi-Holt approach to automatic data editing of coded (categorical) data compares favorably with the more recently proposed techniques. However, his paper did not contain rigorous proofs of the results arising from Fellegi and Holt, nor did it investigate the appropriateness of the aforementioned objective function. The present paper contains the necessary proofs, identifies the relationship between a statistical error model and the objective function, and discusses the requirement that solutions be independent of the representation of the edits.

2. CONNECTIONS TO MAXIMUM POSTERIOR PROBABILITY

The error detection techniques discussed are deterministic; that is, records failing the edits (constraints) are assuredly in error. Symbolically, if $F_i(x)$ indicates that the i^{th} field of the record x is in error, then an edit e is deterministic if and only if, whenever x fails the edit, it follows that at least one of the fields is in error. If the edit e specifies a failure region and the statement that x fails e is written as $x \in e$, then

$$P\{\cup F_i(x) \mid x \in e\} = 1 ,$$

where $\cup F_i(x)$ is defined to mean that some field of x is in error. Hence, the statistical interest in the problem is shifted from the determination of erroneous records to the determination of the most likely cause for the failure of each such erroneous record (localization

of the error, often called the "fields to impute problem"). In one formulation, the problem becomes the identification of the subset of fields associated with the maximum posterior probability that errors occur in exactly these fields, conditioned on the fact that the record has failed specified edits. Again symbolically, if S is the collection of all subsets of the field indices $\{1, \dots, n\}$, if $F^S(x)$ is defined to mean that exactly the fields indexed by $i \in s \subset S$ are in error, and if $T_m(x)$ is defined to mean that $x \in e_{ji}$ for $i = 1, \dots, m$, then the fields to impute problem becomes

$$\max_{s \subset S} P\{F^S(x) \mid T_m(x)\} , \quad (1)$$

subject to the conditions that $x + y$ is an acceptable record and $y_i = 0$ for $i \notin s$. (Clearly, in this formulation y is the "correction factor" to the record x .)

It is important to note that for the solution to be meaningful, it must be independent of the representation of the edits; that is, if $\{e_k\}$, $k \in C'$, is another collection of edits that specifies exactly the same locus of points as $\{e_j\}$, $j \in C$, that is

$$\bigcup_{C'} e_k = \bigcup_C e_j ,$$

and if the record x fails exactly edits $e_{k1}, \dots, e_{kw}$ of C' , written as $T'_w(x)$, then it must follow that the s that maximizes $P\{F^S(x) \mid T_m(x)\}$ must also maximize $P\{F^S(x) \mid T'_w(x)\}$ and conversely. Generally, the maximum posterior probability is constrained to be invariant of representation by constraining the allowable collections of edits. This point is further discussed in Sects. 5 and 6.

Generally, due to insufficient information about the necessary probability distributions, no reasonable method is known to directly evaluate the conditional probabilities involved in Eq. (1) or to rigorously transform them into a more convenient format. Even were the information available, the required calculations would often be impractical. However, Naus, Johnson, and Montalvo¹ have demonstrated a limited result that suggests the direction of a simplifying transformation. Specifically, let $\{e_j\}$, $j = 1, \dots, m$ be a set of edits failed by the record x . Let D_i be a set of fields such that for each edit $e \in \{e_j\}$, there exists a field $k \in D_i$ that is restricted by the edit e . Let $B_i = F^{D_i}(x)$. Under conditions similar to those derived by Naus, Johnson, and Montalvo,^{*} it can be shown that

$$P\{F^S(x) \mid T_m(x)\} = P\{F^S(x) \mid \bigcup_J B_i\},$$

where the union is taken over all possible sets satisfying the property given above. The latter conditional probability can be rewritten as

$$P\{F^S(x) \mid \bigcup_J B_i\} = \frac{P\{F^S(x) \cap (\bigcup_J B_i)\}}{P\{\bigcup_J B_i\}}.$$

^{*} Assume $P\{F^S(x) \mid T_m(x)\} = P\{T_m(x) \mid \bigcup_J B_i\}$. Then $P\{F^S(x) \mid T_m(x)\}$

$$= \frac{P\{T_m(x) \mid F^S(x)\} P\{F^S(x)\}}{P\{T_m(x)\}} = \frac{P\{T_m(x) \mid F^S(x)\} P\{F^S(x)\}}{P\{T_m(x) \cap (\bigcup_J B_i)\}}$$

$$= \frac{P\{T_m(x) \mid F^S(x)\} P\{F^S(x)\}}{P\{T_m(x) \mid \bigcup_J B_i\} P\{\bigcup_J B_i\}} = P\{F^S(x) \mid \bigcup_J B_i\}.$$

From the definition of the term $F^S(x)$, it follows that the maximization of Eq. (1) can be identified with

$$\max_{s \in S} P\{F^S(x)\} ,$$

subject to the following conditions: $x + y$ is an acceptable record and $y_i = 0$ for $i \notin s$. However, even this reduction of the problem leaves unanswered the question of the best way to determine the prior probabilities (which is equivalent to specifying an "error model").

In any case, given that the prior probabilities can be somehow estimated, two additional desiderata should be considered. First, the solution to the fields to impute problem should be somewhat robust with respect to error models, and second, the algorithm used to solve the problem must be efficient. Whether or not some degree of robustness will result remains to be seen. The emphasis on efficiency is placed in better perspective when one realizes, for example, that editing the Canadian census involves some 40×10^6 records and that an error rate of 1 to 10% is typical;⁶ therefore, the fields to impute problem must be solved on the order of 4×10^6 times. Even for one of the simplest cases, where the probability p_i of an error in field i is independent of the magnitude of the error and is also independent of the probability p_j of an error in field j , $j \neq i$, the fields to impute problem is a difficult nonlinear constrained optimization problem. In fact, if the conditional probability $P\{F^S \mid T_m\}$ is approximated by $P\{F^S \mid \bigcup_J B_i\}$, as suggested by Naus, Johnson, and Montalvo,¹ the problem can be written as

$$\max p_s, \quad p_s = \prod_{i \in s} p_i \left[\prod_{i \notin s} (1 - p_i) \right] ,$$

subject to the conditions that $x + y$ is an acceptable record and $y_i = 0$, $i \notin s$.

This problem can be transformed into a more convenient form by the observation that $\prod_{i \in s} p_i \left[\prod_{i \notin s} (1 - p_i) \right] = \prod_{i=1}^n (1 - p_i) \prod_{i \in s} \frac{p_i}{(1 - p_i)}$. Now for $p_i \leq 1/2$, set $c_i = -\ln p_i + \ln (1 - p_i)$ and for $1 > 1/2$ set $c_i = \ln p_i - \ln (1 - p_i)$. Then the problem can be written

$$\min \sum c_i \delta(y_i) , \quad (2)$$

subject to the following conditions: $x + y$ is an acceptable record and

$$\delta(y_i) = \begin{cases} 0, & y_i = 0 \\ 1, & y_i \neq 0 \end{cases} .$$

The case where all the probabilities are equal, $p_i = p_j$, is exactly the point of departure for Fellegi and Holt. To be sure, one suspects that they were aware of a simple tree enumeration scheme to solve the problem, but realized that it was highly inefficient and therefore developed an alternative algorithm. Central to the satisfactory operation of their algorithm is the need to express the edits in normal form and to generate a sufficient set of (implied) edits. This dependence on the implied edits is the strength of the Fellegi-Holt algorithm as well as a potential indictment against it; the implied edits make each fields to impute problem more tractable, but the generation of these edits could in certain cases still be a major task.

3. THEORETICAL DETAILS OF THE FELLEGI-HOLT APPROACH

This section rigorously presents the central ideas of the Fellegi-Holt approach to automatic data editing, including the usefulness of the sufficient class of implied edits to the fields to impute problem. *The theory is presented exclusively for coded data.*

Definition 1. The (univariately) feasible values for field i are A_i .

Definition 2. Let each record of the data set have n fields; then the data space is defined to be the Cartesian product

$$\prod_{i=1}^n A_i .$$

Definition 3. An edit is a consistency check that defines a failure region. An edit e is in normal form if and only if it can be expressed as a Cartesian product, that is,

$$e = \prod_{i=1}^n \pi_i(e) ,$$

where $\pi_i()$ is the projection on the i^{th} field. Often an edit e will be written as

$$e = \prod_{i=1}^n A_i^r$$

where clearly $A_i^r = \pi_i(e)$.

Definition 4. A normal-form edit e is *essential* if and only if $A_i^r = A_i$ for at least one of the fields. Each such field is called an essential field.

Definition 5. A field i *enters* the normal-form edit e if and only if it is not an essential field of e . (The fields that enter an edit are precisely those fields that in Naus's terminology are involved in the edit.)

Definition 6. Let $\{e_i\}$ be a collection of edits. Then an edit e is an *implied edit* (by collection $\{e_i\}$) if and only if $e \subset \cup e_i$, where the union and inclusion are taken in the point-set sense.

Example 1

Consider a data space $A_1 \times \dots \times A_6$ where

$$A_1 = \{0, 1\} \quad A_4 = \{0, 1, 2, 3\}$$

$$A_2 = \{0, 1, 2\} \quad A_5 = \{0, 1, 2\}$$

$$A_3 = \{0, 1\} \quad A_6 = \{0, 1, 2, 3\}$$

and explicit edits

$$e_1 = A_1 \times \{0, 1\} \times \{0\} \times A_4 \times \{0, 1\} \times A_6$$

$$e_2 = \{1\} \times A_2 \times \{1\} \times \{0, 1\} \times A_5 \times \{2, 3\}$$

$$e_3 = \{0\} \times \{1, 2\} \times A_3 \times \{1, 2, 3\} \times A_5 \times A_6$$

$$e_4 = A_1 \times \{0, 2\} \times A_3 \times A_4 \times A_5 \times \{0, 1\}$$

$$e_5 = \{1\} \times A_2 \times A_3 \times \{0\} \times \{1, 2\} \times A_6$$

In this example, fields 1, 4, and 6 are essential fields of edit e_1 .

Moreover, it can easily be verified, for example, that $e_6 = A_1 \times \{1, 2\} \times \{1\} \times \{1\} \times A_5 \times \{2, 3\}$ is an implied edit.

The natural question about the expressability of any edit in normal form is answered by lemma 1.

Lemma 1. Let each A_i be bounded; then every edit can be expressed as the finite union of normal-form edits.

Proof. Each point

$$x \in \prod_{i=1}^n A_i$$

is in normal form. There are only finitely many such points.

Henceforth, only normal-form edits will be considered. As alluded to earlier, certain implied edits are critical to the functioning of the Fellegi-Holt approach to data editing. Lemma 2 provides a means for constructing these. (The statement of the lemma is essentially unmodified from that of Fellegi and Holt.)

Lemma 2. Let

$$e_r = \prod_{i=1}^n A_i^r, \quad r \in S$$

be edits. Assume that

$$\emptyset \neq A_i^* = \bigcap_{r \in S} A_i^r \text{ for } i \neq j,$$

and that $\emptyset = A_j^r$, $r \in S$, and hence that

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

Then

$$e^* = \prod_{i=1}^n A_i^*$$

is an implied edit.

Proof. It suffices to show that

$$\prod_{i=1}^n A_i^* \subset \left[\bigcup_{r \in S} \prod_{i=1}^n A_i^r \right].$$

Assume that $a = (a_1, \dots, a_n) \in e^*$. Hence it follows for some $t \in S$ that $a_j \in A_j^t$. Then for each $i \in (1, \dots, n)$, $a_i \in A_i^t$.

Definition 7. The field indexed by j in lemma 2 is the *generating field* for e^A .

Example 2

The edit e_6 (of Example 1) can be generated from edits e_2 and e_3 , with field 1 the generating field.

Ultimately, it will be necessary to know that the method of lemma 2 is adequate to generate a sufficient collection of implied edits to solve the fields to impute problem [Eq. (2)]. (The connection between the fields to impute problem and implied edits is specified later in theorem 4.) Lemma 3, which is a correction of theorem 2 of Fellegi and Holt, indicates that "large" implied edits can be generated.

Lemma 3. Let e be any implied edit. Then application of the method of lemma 2 on the explicit edits and repeated application of the method on the resultant edits will yield an edit e^* that satisfies $e \subset e^*$.

Proof. Consider any implied edit $e = A_1^r \times \dots \times A_n^r$. The edit can be written as

$$e = \bigcup_{a \in A_1^r} (a \times A_2^r \times \dots \times A_n^r).$$

Hence, by induction, it suffices to prove that an edit can be generated that covers each $A^* = a_1 \times \dots \times a_{n-1} \times A_n^r \subset A_1^r \times \dots \times A_n^r$. Choose a minimal set S of explicit edits that satisfies

$$A^* \subset \bigcup_{r \in S} e_r.$$

The implied edit e^* generated by the set S of edits with generating field n satisfies $A^* \subset e^*$. Specifically, for each $a_n^j \in A_n^r$, there exists $e_{r(j)} \in S$ satisfying $(a_1, \dots, a_{n-1}, a_n^j) \in e_{r(j)}$, and by minimality of S , each edit $e_{r(j)} \in S$ contains some element $(a_1, \dots, a_{n-1}, a_n^j)$.

Later it will be shown that an appropriate subset of the collection of all explicit and implied edits exists that enables one to reformulate the fields to impute problem [Eq. (2)] as a set-covering problem with no explicit mention of the acceptance region. The framework for this result is contained in lemma 4, corollaries 1 and 2, and culminates in theorem 1.

Definition 8. Let E be a collection of explicit edits and C be a subset of the indices $\{1, \dots, n\}$. Then with respect to E the collection Ω_C of edits (both explicit and implied) is defined to be the set of all possible edits satisfying

$$\Omega_C = \{e_r = \prod_{i=1}^n A_i^r \mid A_j^r = A_j, j \in C, e_r \subset \bigcup_E e\}.$$

Note: For any well-ordered sequence $C_1 \subset C_2 \subset \dots \subset C_n$, the relation $\Omega_{C_n} \subset \dots \subset \Omega_{C_1}$ obtains, and if C is empty, then Ω_C is taken to be the collection of all edits.

Example 3

The edit e_5 (of Example 1) satisfies $e_5 \in \Omega_{\{2,3,6\}}$.

Lemma 4 (Fellegi and Holt). Let $C_{k-1} \subset C_k$, $C_k \sim C_{k-1} = \{i \mid i \in C_{k-1}, i \notin C_k\}$ and assume that for all edits $e \in \Omega_{C_k}$, the relationship $a^0 \notin e$ obtains. Then the point a^0 can be modified by changing at most coordinate $i \in C_k \sim C_{k-1}$ so that the resultant point $\hat{a}$ satisfies the following relationship: $\forall e \in \Omega_{C_{k-1}}, \hat{a} \notin e$. Symbolically, if $\forall e \in \Omega_{C_k}, a^0 \notin e$, then there exists a point $\hat{a}$ satisfying $\Pi_i(\hat{a}) = \Pi_i(a^0)$ for $i \notin C_k \sim C_{k-1}$ and

$$\forall e \in \Omega_{C_{k-1}}, \hat{a} \notin e.$$

Proof. Without loss of generality, let $C_k = \{1, \dots, k\}$. Now assume to the contrary that for any value a_k of the k^{th} coordinate field, the point $(a_1^0, \dots, a_{k-1}^0, a_k, a_{k+1}^0, \dots, a_n^0)$ fails an edit of $\Omega_{C_{k-1}}$. From the definition of C_{k-1} , it follows that, for any values $a_1, \dots, a_k$ of the first k coordinates, the point $(a_1, \dots, a_k, a_{k+1}^0, \dots, a_n^0)$ fails an edit of $\Omega_{C_{k-1}}$ and hence fails an edit of Ω_{C_k} , which is a clear contradiction.

Corollary 1. Assume that $\forall e \notin \Omega_{C_k}, a^0 \notin e$. Then there exists a point a satisfying $\Pi_i(\hat{a}) = \Pi_i(a^0)$ for $i \notin C_k$ and $\forall e \in \Omega_{\emptyset}, \hat{a} \notin e$.

Proof. Repeated application of lemma 4.

At this point it is perhaps worthwhile to underscore the necessity of the full set of edits considered in lemma 4. The crucial step is the implication $[\forall \{a_1, \dots, a_k\}, \exists e \in \Omega_{C_{k-1}} : a = (a_1, \dots, a_k, a_{k+1}^0, \dots, a_n^0) \in e] \Rightarrow [\exists e \in \Omega_{C_k}, a \in e]$; it is crucial to recognize the existence of the edit belonging to Ω_{C_k} .

Definition 7. A set S of edits is *sufficient* if and only if, for each data element x and each maximal (with regard to inclusion) edit (either explicit or implied) with fields $i_1, \dots, i_k$ essential and $x \in e$, there exists an edit $\hat{e} \in S$ satisfying $x \in \hat{e}$, and fields $i_1, \dots, i_k$ are essential.

Corollary 2. Let $\Omega_{C'} = \{e \in \Omega_C \mid e \text{ is a maximal edit}\}$. Replace Ω_{C_k} and $\Omega_{C_{k-1}}$ in lemma 4 and corollary 1 with $\Omega_{C'_k}$ and $\Omega_{C'_{k-1}}$ respectively. Then the conclusion still holds.

Proof. Obvious.

Corollary 3. Let S be a sufficient set of edits and $\Omega_{C''} = \{e \in \Omega_C \text{ and } e \in S\}$. Replace Ω_{C_k} and $\Omega_{C_{k-1}}$ in lemma 4 and corollary 1 with $\Omega_{C''_k}$ and $\Omega_{C''_{k-1}}$ respectively. Then the conclusion still holds.

Proof. Obvious.

Theorem 4 relates the edits to the fields to impute problem. It differs from corollary 2 of the Fellegi-Holt paper insofar as it requires a sufficient set of edits and the original Fellegi-Holt work required essentially all edits. From the perspective of the formal proof, the strengthening of the result is inconsequential, but from the perspective of implementability, the difference is substantial.⁵

Theorem 4. Let $\Omega_n = \phi$, S be a sufficient set of edits, $c_i \geq 0$, and furthermore let E be the set of edits of S failed by the point x^0 . Let I be a (weighted) minimal set of field indices satisfying the property that if $e \in E$, then there exists $i \in I$ such that the field i enters the edit e , that is, $\forall e \in E, \exists i, i \in I$, which enters e , and $\sum_{i \in I} c_i$ is minimized. Then this set I is precisely a solution (possibly nonunique) to the fields to impute problem insofar as a point x exists that passes

all edits and satisfies $\hat{x}_i = x_i^0$ for $i \notin I$, $\hat{x}_i \neq x_i^0$ for $i \in I$, and furthermore that no smaller (weighted) number of changes can be made to make the record x^0 acceptable.

Proof. From corollary 1, for each edit $e \in E$, at least one of the fields that enters the edit e must be changed to make the record x^0 pass that edit. The collection I of fields minimally satisfies this criterion for each of the edits failed, so that at the least each of these fields must be changed. Were some additional change necessary, it would follow that there exists an edit e for which all the fields $i \in I$ are essential and $x^0 \in e$. This contradicts the choice of I .

Remark. Again it is important to note the necessity of the sufficient set of edits.

4. THE SET-COVERING PROBLEM

Example 4

Consider again the data space of Example 1 and the five edits.

$$e_1 = A_1 \times \{0, 1\} \times \{0\} \times A_4 \times \{0, 1\} \times A_6$$

$$e_2 = \{1\} \times A_2 \times \{1\} \times \{0, 1\} \times A_5 \times \{2, 3\}$$

$$e_3 = \{0\} \times \{1, 2\} \times A_3 \times \{1, 2, 3\} \times A_5 \times A_6$$

$$e_4 = A_1 \times \{0, 2\} \times A_3 \times A_4 \times A_5 \times \{0, 1\}$$

$$e_5 = \{1\} \times A_2 \times A_3 \times \{0\} \times \{1, 2\} \times A_6$$

Let the current record be $(1, 0, 0, 0, 1, 0)$. This record fails edits e_1 , e_4 , and e_5 . Accordingly, were these five explicit edits sufficient to determine the fields to impute, one would solve the following set-covering problem to identify these fields:

$$\min \sum w_i$$

$$\text{subject to } \begin{vmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{vmatrix} w \geq \begin{vmatrix} 1 \\ 1 \\ 1 \end{vmatrix}, \quad w_i = 0, 1.$$

This problem corresponds to finding the minimal collection of fields such that for each edit failed, at least one field enters the edit. In particular, the first row of the matrix indicates that the first field does not enter the edit e_1 , that the second and third do, and that with the exception of field 5 the remaining ones do not. One minimal solution is $w = (1, 1, 0, 0, 0, 0)$, which would suggest in the case of a sufficient set of edits that there would be a change of the entries in only the first two fields, ensuring that the record passes edits e_1 , e_4 , and e_5 . In fact, no such change leads to the satisfaction of these edits; since $\{e_1, \dots, e_5\}$ is not a sufficient set of edits, it develops that any such modified record fails either edit e_1 or e_4 . Loosely speaking, this indicates that there is an implied edit that contains the original record and for which the first two fields are essential; this observation provides the basis for an algorithm that is an alternative to that of Fellegi and Holt because it does not require the preliminary determination of implied edits.⁷ On the other hand, one could determine the following sufficient collection of edits:

$$\{e_1, \dots, e_5\}$$

$$e_6 = A_1 \times \{1, 2\} \times \{1\} \times \{1\} \times A_5 \times \{2, 3\}$$

$$e_7 = A_1 \times A_2 \times \{0\} \times A_4 \times \{0, 1\} \times \{0, 1\}$$

$$e_8 = A_1 \times \{1\} \times A_3 \times \{1\} \times \{0, 1\} \times \{2, 3\}$$

$$\begin{aligned}
e_9 &= A_1 \times \{2\} \times \{1\} \times \{1\} \times A_5 \times A_6 \\
e_{10} &= \{0\} \times A_2 \times \{0\} \times \{1, 2, 3\} \times \{0, 1\} \times A_6 \\
e_{11} &= \{0\} \times A_2 \times A_3 \times \{1, 2, 3\} \times A_5 \times \{0, 1\} \\
e_{12} &= \{1\} \times A_2 \times \{0\} \times \{0\} \times A_5 \times \{0, 1\} \\
e_{13} &= \{1\} \times \{0, 1\} \times A_3 \times \{0, 1\} \times \{0, 1\} \times \{2, 3\} \\
e_{14} &= \{1\} \times \{0, 1\} \times A_3 \times \{0\} \times A_5 \times \{2, 3\} \\
e_{15} &= \{1\} \times \{0, 1\} \times \{0\} \times \{0\} \times A_5 \times A_6 \\
e_{16} &= \{1\} \times \{0, 2\} \times \{1\} \times \{0, 1\} \times A_5 \times A_6 \\
e_{17} &= \{1\} \times \{0\} \times A_3 \times \{0\} \times A_5 \times A_6 \\
e_{18} &= \{1\} \times \{0\} \times A_3 \times \{0, 1\} \times \{0, 1\} \times A_6 .
\end{aligned}$$

In terms of this collection, the record (1, 0, 0, 0, 1, 0) fails edits $e_1, e_4, e_5, e_7, e_{12}, e_{15}, e_{17},$ and e_{18} . Given unit costs, the fields to impute problem becomes

$$\begin{array}{rcl}
& \min \sum w_i \\
\text{subject to} & \begin{vmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{vmatrix} & w \geq \begin{vmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{vmatrix}, \quad w_i = 0, 1 .
\end{array}$$

One solution is $w_1 = w_2 = w_3 = 1$, and $w_4 = w_5 = w_6 = 0$. For example, the record (0, 1, 1, 0, 1, 0) passes all edits.

In general, with use of the Fellegi-Holt results, the fields to impute problem (2) can be written as a set-covering problem (Eqs. 3 and 4):

$$\min \sum c_i w_i, \tag{3}$$

subject to

$$Mw \geq 1, \quad w_i = 0, 1, \quad (4)$$

and M is the failed-edit matrix (local to the erroneous record x) with rows indexed by the failed edits, columns by the fields, 1's corresponding to fields that enter, and 0's to essential fields. *Moreover, it is not hard to see and is worthwhile to note that (if one generates a collection of edits as in corollary 3) the solution set is invariant of the original representation of the explicit edits; any two representations that cover the same locus in space yield the same solution set.*

5. MODIFICATIONS OF THE WEIGHTING

In a paper by Freund and Hartley,⁸ it is suggested that the potential corrections be weighted by the "number of involvements in unsatisfied consistency checks." In fact, in the case considered, they found that such weighting yielded better results than otherwise obtained. Although Freund and Hartley investigated weighting exclusively for continuous data and as part of a least squares imputation process, weighting is not unreasonable for the fields to impute problem for coded data and is equivalent to modifying the objective function [Eq. (2)], making it dependent on the edits failed. (The hope would be that the improvement observed by Freund and Hartley would hold for this case also.) Care need only be exercised that a disjoint sufficient collection of edits be used for the sufficient collection;* if not, the objective function may

*The collection of all possible edits or all maximal edits would also be acceptable (although in general the relative weighting might be slightly different in these cases), but the disjoint sufficient collection seems the most natural and least unwieldy.

not be independent of the original edit representation. Further details are developed below.

Definition 8. A sufficient collection E of edits is a disjoint sufficient collection if and only if whenever $e_1 \cap e_2 \neq \phi$, $e_1, e_2 \in E$, the essential fields of the intersection $e_1 \cap e_2$ are a proper subset of the essential fields of both e_1 and e_2 .

Definition 9. Consider a field i and a set of edits E . Define $n_i(x)$ to be the reciprocal of the number of involvements of the field i in edits $\hat{E} \subset E$ failed by the record x ,

$$[\sum_{e \in \hat{E}} f_i(e)]^{-1},$$

where

$$f_i(e) = \begin{cases} 1 & \text{if field } i \text{ enters edit } e \\ 0 & \text{otherwise} \end{cases}$$

(The problem of unboundedness of the reciprocal whenever

$$\sum_{e \in \hat{E}} f_i(e) = 0$$

can be handled by defining it to be some arbitrary large number, or by omitting the corresponding variable from the weighted objective function.

Example 5

Consider the edits of Example 4 and the same record $x = (1, 0, 0, 0, 1, 0)$. If only the original explicit edits are used to determine the objective function, then the weighting of Eq. (3) becomes

$$x_1 + \frac{1}{2} x_2 + x_3 + x_4 + \frac{1}{2} x_5 + x_6.$$

In the case of Example 4, the sufficient set of edits generated is fortuitously a disjoint sufficient set and also a set of maximal edits. For this set the objective function becomes

$$\frac{1}{5} x_1 + \frac{1}{5} x_2 + \frac{1}{4} x_3 + \frac{1}{5} x_4 + \frac{1}{4} x_5 + \frac{1}{3} x_6 .$$

Example 6

Let the data space be $\{0, 1\} \times \{0, 1\} \times \{0, 1\}$, and let edits be defined as per five different cases.

Case 1	Case 2	Case 3
$e_1 = \{0\} \times \{0\} \times \{0\}$	$e_1 = \{0, 1\} \times \{0\} \times \{0\}$	$e_1 = \{0\} \times \{0, 1\} \times \{0\}$
$e_2 = \{0\} \times \{1\} \times \{0\}$	$e_2 = \{0\} \times \{1\} \times \{0\}$	$e_2 = \{1\} \times \{0\} \times \{0\}$
$e_3 = \{1\} \times \{0\} \times \{0\}$		

Case 4	Case 5
$e_1 = \{0, 1\} \times \{0\} \times \{0\}$	$e_1 = \{0\} \times \{0\} \times \{0\}$
$e_2 = \{0\} \times \{0, 1\} \times \{0\}$	$e_2 = \{0\} \times \{1\} \times \{0\}$
	$e_3 = \{1\} \times \{0\} \times \{0\}$
	$e_4 = \{0, 1\} \times \{0\} \times \{0\}$
	$e_5 = \{0\} \times \{0, 1\} \times \{0\}$

The locus covered by the edits is identical in each of the five cases. Now consider the record $(0, 0, 0)$. The objective function for each of the five cases becomes

Case 1: $x_1 + x_2 + x_3$

Case 2: $x_2 + x_3$

Case 3: $x_1 + x_3$

Case 4: $x_1 + x_2 + \frac{1}{2} x_3$

Case 5: $\frac{1}{2} x_1 + \frac{1}{2} x_2 + \frac{1}{3} x_3$.

Symmetry suggests that cases 2 and 3 result in inappropriate objective functions (unless additional overriding information is available on the error process). Similarly, justification can be presented (over and above the invariance to the original edit representation) for preference for the case 4 objective function rather than the case 1 function.⁹

In this example, case 4 is the disjoint sufficient collection of edits and is identical to the collection of all maximal edits, and case 5 is the collection of all possible edits.

6. DETAILS OF INVARIANCE

Lemma 5. Let

$$L \subset \prod_{i=1}^n A_i$$

be a locus of points. Then the collection of all edits covering L and the collection of all maximal edits covering L is only dependent on L and not on the original covering of L by explicit edits.

Proof. Obvious.

Theorem 5. Let

$$L \subset \prod_{i=1}^n A_i$$

be a locus of points. Then there are possibly nonunique disjoint sufficient collections of edits that cover L . However, if E_1 and E_2 are two such disjoint sufficient collections of edits and x is any record, then for any field i , $n_i(x)$ as calculated from E_1 equals $n_i(x)$ calculated from E_2 .

Proof. Step 1, possible nonuniqueness.

Let $A = \{0, 1, 2\} \times \{0, 1\} \times \{0, 1\}$. Let $E_1 = \{(\{1\} \times \{0, 1\} \times \{0\}), (\{2\} \times \{0, 1\} \times \{0\})\}$ and $E_2 = \{\{1, 2\} \times \{0, 1\} \times \{0\}\}$. It is a straightforward check that E_1 and E_2 are both disjoint sufficient collections.

Step 2, field involvements well defined.

For any disjoint-sufficient collection E of edits, record x , and fields $i_1, \dots, i_k$, whether or not an edit $e \in E$ exists for which exactly these fields enter and $x \in e$ depends solely on the existence of a maximal edit $\hat{e}$ with exactly the same fields entering and satisfying $e \subset \hat{e}$. In any case, from the definition of a disjoint collection, there exists at most one such edit $e \in E$. Thus the invariance follows from lemma 5.

It should be noted that two distinct types of invariance have been discussed. The first presumes a fixed (given) objective function and is the invariance of the solution to the fields to impute problem with respect to the initial specification of the explicit edits for the aforementioned objective function. It has been shown that for this invariance to hold and for the solution of the fields to impute problem to be obtainable, the original explicit edits need only to be reexpressed as a sufficient collection of edits. The second type of invariance is the invariance of the objective function derived in terms of involvements.

Again, the invariance is to be with respect to the initial specification of the explicit edits. In this case it was shown that a disjoint sufficient collection of edits needed to be generated from the original edits to ensure invariance. Hence, a reexpression of the original explicit edits as a disjoint sufficient collection would be adequate for both types of invariance.

7. COMPUTATIONAL DETAILS FOR DISJOINT SUFFICIENT COLLECTIONS

At this point, a natural question is the feasibility and difficulty of identifying a disjoint sufficient collection of edits. The answer is contained principally in Liepins,⁵ where the modification to algorithm 1 is expected to result in a reasonably efficient generation of a sufficient collection of edits.* Straightforward manipulation of the resultant sufficient collection of edits leads to a disjoint sufficient collection.

The following algorithm reexpresses a sufficient collection of edits as a disjoint sufficient collection.

1. Group the edits by essential field combinations.
2. For each essential field combination that has more than one edit associated to it, follow these steps:
 - a. Consolidate edits. Within each field combination, if any set of edits is identical except for the values in one field, use that field as the generating field as in lemma 2 and replace the set of generating edits with the one newly generated edit.
 - b. Order the edits.

* Recent work seems to be substantiating the expectations.

c. For $i = 1, \dots$ and $j = i + 1, \dots$, form the complements

$$e_i \sim e_j.$$

3. Form the directed "field combination" graph as follows: Let the nodes be the field combinations. A field combination $ab \dots c$ is connected to another $de \dots f$ by a directed edge pointing from $ab \dots c$ to $de \dots f$ if and only if $\{d, e, \dots, f\} \subset \{a, b, \dots, c\}$ where the inclusion is intended as setwise inclusion of field indices.
4. For any two nodes $ab \dots c$ and $de \dots f$ connected by a directed edge pointing from $ab \dots c$ to $de \dots f$ and edits e_i associated with field combination $ab \dots c$, e_j associated with field combination $de \dots f$, form the complements $e_i \sim e_j$.

It is easily verified that this algorithm accomplishes its desired purpose and that it is finite (there are only finitely many points in the space A). No claim of optimality is made for this algorithm; in fact, straightforward variants with improved performance are readily identified. However, the effort to develop an "optimal" such algorithm will not be made until the Freund-Hartley weighting can be shown to improve the results of discrete data error localization.

Example 7

Consider the edits of Example 4. They were determined using algorithm 1 of Liepins⁵ and hence form a sufficient collection. The grouping by essential field combinations is given below.

Essential fields	Edits
1, 4, 6	e_1
2, 5	e_2, e_{12}
3, 5, 6	e_3, e_{17}
1, 3, 4, 5	e_4
2, 3, 6	e_5
1, 5	e_6
1, 2, 4	e_7
1, 3	e_8
1, 5, 6	e_9
2, 6	e_{10}
2, 3, 5	e_{11}
3	e_{13}
3, 5	e_{14}
5, 6	e_{15}, e_{16}
3, 6	e_{18}

Only three field combinations, 2, 5; 3, 5, 6; and 5, 6 contain more than one edit. None of these has a nonempty intersection of edits, and hence step 2 leaves the edits unchanged. At step 4, edit 18, for example, must be modified by edits e_5 , e_3 , and e_{17} . The result is the edit $\{1\} \times \{0\} \times A_3 \times 1 \times \{0,1\} \times A_6$. The disjoint sufficient collection requires changes only in field combinations 3; 1, 5; 3, 5; 5, 6; and 3, 6. To these field combinations are associated the following edits.

Essential fields	Edits
3	$\emptyset$
1, 5	$A_1 \times \{1\} \times \{1\} \times \{1\} \times A_5 \times \{2,3\}$
3, 5	$\{1\} \times \{1\} \times A_3 \times \{0\} \times A_5 \times \{2,3\}$
5, 6	$\left\{ \begin{array}{l} \{1\} \times \{1\} \times \{0\} \times \{0\} \times A_5 \times A_6 \\ \{1\} \times \{0\} \times \{1\} \times \{1\} \times A_5 \times A_6 \\ \{1\} \times \{2\} \times \{1\} \times \{0\} \times A_5 \times A_6 \end{array} \right.$
3, 6	$\{1\} \times \{0\} \times A_3 \times \{1\} \times \{0,1\} \times A_6$

8. CONCLUSIONS

- For purposes of solving the fields to impute problem, a sufficient collection of edits (as defined in definition 7) is adequate and can be generated by algorithm 1 of Liepins.⁵ The solution to the fields to impute problem is invariant of the particular sufficient collection identified.
- The weighting by number of involvements (Sect. 6) appears to be a reasonable way to modify the objective function of the fields to impute problem. However, care must be taken that the assigned function be independent of the original representation of the edits. Any disjoint-sufficient collection of edits (definition 8) ensures that the invariance obtains (theorem 5).
- In the case of independence of errors by fields and magnitude and the satisfaction of appropriate symmetry conditions,¹ there is a direct correspondence between determining the maximum posterior probability of error combinations and minimizing a linear objective function of the fields to impute problem.

REFERENCES

1. J. I. Naus, T. G. Johnson, and R. Montalvo, "A Probabilistic Model for Identifying Errors in Data Editing," *J. Am. Stat. Assoc.* 67(340): 943-50 (1972).
2. I. P. Fellegi and D. Holt, "A Systematic Approach to Automatic Edit and Imputation," *J. Am. Stat. Assoc.* 71(353): 17-35 (1976).
3. G. Sande, *An Algorithm for the Fields to Impute Problems of Numerical and Coded Data*, unpublished paper.
4. P. G. McKeown, *A Branch-and-Bound Algorithm for Solving Fixed Charge Problems*, University of Georgia, unpublished paper.
5. G. E. Liepins, *Refinements to the Boolean Approach to Automatic Data Editing*, to be published as an ORNL report.
6. J. I. Naus, *Data Quality Control and Editing*, Marcel Dekker, Inc., New York, 1975.
7. R. S. Garfinkel, *An Algorithm for Optimal Imputation of Erroneous Data*, unpublished paper, 1979.
8. R. J. Freund and H. O. Hartley, "A Procedure for Automatic Data Editing," *J. Am. Stat. Assoc.* 62(318): 341-52 (1967).
9. E. T. Jaynes, "Prior Probabilities," *IEEE Trans. Systems Sci. Cybern.* SSC-4 (3): 227-41 (September 1968).

INTERNAL DISTRIBUTION

- 1-24. A. S. Loeb1
- 25-26. Central Research Library
- 27. Document Reference Section
- 28-29. Laboratory Records
- 30. Laboratory Records - RC
- 31. ORNL Patent Section

EXTERNAL DISTRIBUTION

- 32. Office of Assistant Manager for Energy Research and Development,
DOE/ORO
- 33-47. G. E. Liepins, c/o University of Georgia, Dept. of Quantitative
Analysis, Athens, GA 30602
- 48. Johannes Wetzig, Arbeitsgruppe Hochenergiephysik, Sektion
Physick, Karl-Marx-Universitat, DDR-701 Leipzig, Karl-Marx-Platz
- 49-231. Given to distribution as shown in TIC 4500 under UC-32 Category