AN INTEGRATED APPROACH TO DATA EDITING

Gunar E. Liepins and David J. Pack
Oak Ridge National Laboratory
Union Carbide Corporation

ABSTRACT

The problem considered is to determine how to localize error in particular fields of categorical records from large data sets that fail to fall in a deterministically known constrained space. The so-called minimal fields to impute (MWFI) error localization is to find the set S of the records's m fields which minimizes

$$\sum_{i=1}^{m} c_i \delta(\varepsilon_i) ,$$

where ε_i is the field i change, $\delta(\varepsilon_i) = 0$ if $\varepsilon_i = 0$ and 1 otherwise, and the c_i weights are decreasing functions of the probability of error in field i, and the minimization is subject to the constraint that the changed record be acceptable. Historically the c_i weights have been assumed equal for all fields - effectively representing uniform prior probabilities of error. Optimally these weights should be probabilities of error given the particular erroneous error, i.e., posterior probabilities. This paper reports on a simulation study of a compromise in which the weights are related to the number of active failed edits and/or the number of active passed edits for a field.

I. INTRODUCTION

Automatic data editing is the computerized identification, localization, and correction of data errors in situations where records from a data base must theoretically fall in a deterministically known constrained space. Identification is the process of finding records in the data base containing one or more errors. If an observed record $\underline{y} = (y_1, y_2, \ldots, y_m)$ of m fields fails a known constraint, for example,

$$\sum_{i=1}^{m} b_i y_i > d ,$$

the record is deterministically known to be in error, i.e, the identification is deterministic. Localization is the process of finding the set of fields in the record which actually is in error, while correction is the choice of "true" values for the localized fields ("imputation" is typically used instead of "choice" in the literature). Localization and correction are clearly stochastic in most environments.

The stochastic nature of localization and correction implies a need for a model of the error process. The observed record $\underline{y}$ is the true record $\underline{x}$ plus an error vector $\underline{\varepsilon}$, i.e.,

$$\underline{y} = \underline{x} + \underline{\varepsilon} . \qquad (1.1)$$

Part of the error model is embodied in assumptions about components of the error vector $\underline{\varepsilon}' = (\varepsilon_1, \ldots, \varepsilon_m)$. The remainder of the model focuses on the distribution of the true $\underline{x}$. In principal, knowledge of the error model allows one to solve for the conditional probability

$$\max_{\underline{x}} p(\underline{x}|\underline{y}) \qquad (1.2)$$

which is logically the focus of the localization and correction steps.

2. A STOCHASTIC CONCEPTION OF THE MWFI PROBLEM

The MWFI problem as expressed in the abstract can be derived from error model assumptions which state that

A. $p\{\varepsilon_i \neq 0 | \varepsilon_j \neq 0\} = p\{\varepsilon_i \neq 0\} = p_i$ for $i \neq j$.

B. ε_i has a uniform distribution for all i over the set of feasible values for ε_i.

It follows that the prior probability that exactly the fields $i \varepsilon S$ are in error can be written as

$$J = \prod_{i \varepsilon S} p_i \prod_{i \notin S} (1-p_i) , \qquad (2.1)$$

which is also expressable as

$$J = \prod_{i=1}^{m} (1-p_i) \prod_{i \varepsilon S} p_i \Big/ \prod_{i \varepsilon S} (1-p_i) . \qquad (2.2)$$

One seeks the set S which maximizes J subject to the constraint that the changed record be acceptable. Equivalently one seeks the set S which minimizes

$$\sum_{i \varepsilon S} \log(1-p_i) - \sum_{i \varepsilon S} \log p_i , \qquad (2.3)$$

which is the negative logarithm of (2.2) without the first product since it does not depend on S. But minimization of (2.3) is equivalent to the minimization expressed in the abstract if we define

$$c_i = \log(1-p_i) - \log p_i . \qquad (2.4)$$

Note that (2.4) implies the c_i weights are inversely related to the p_i, the prior probabilities of error in each field. The minimization will thus include in S those fields i with high prior probabilities of error, other things being equal.

3. PROBLEMS WITH MWFI APPLICATION

Problems typically exist with MWFI application as conceived in the previous section, i.e., where the c_i weights are inversely related to prior probabilities of error in fields, and these weights (and, thus, probabilities) are often presumed equal.

It was indicated in Section 1 that the c_i weights should optimally be related to the posterior probabilities of error in fields, which are conditioned on the observed erroneous record. When the c_i weights are presumed equal, problems in MWFI application are compounded. Among these problems is the absence of a unique solution,

i.e., the localization of a unique set of fields as a solution to the error localization process.

4. ALTERNATIVE WEIGHTS

In principal, the problems of Section 3 would be resolved by the use of posterior probabilities of error in fields in determining the c_i weights. One can argue for these probabilities from a conceptual perspective and one can also note that they typically will not be equal for different fields, thus increasing the possibility of a unique solution to the MWFI problem.

In the reality of large data sets (many records, many fields), there are two substantial difficulties in using posterior probabilities. First, one generally possesses insufficient information about the data system to calculate them. This information centers around Eq. (1.1), specifically the so-called "error model". Secondly, given the information to calculate them, one may still find the volume of calculations simply overwhelming.

What seems to be required is some method of using available partial information about the error model to localize field errors in a computationally tractable way which nonetheless is a relatively good approximation to the use of actual posterior probabilities. This paper takes a first step by examining some alternative c_i weights proposed indirectly in a somewhat different context by Freund and Hartley (1967).

The logic of the alternative c_i weights is best introduced via an example. Consider a data base formed by records with 6 fields where a record is one of the members of the Cartesian product $A_1 \times A_2 \times A_3 \; A_4 \times A_5 \times A_6$ and

$$\begin{array}{llll} A_1 = \{0,1\} & & A_4 = \{0,1,2,3\} & \\ A_2 = \{0,1,2\} & & A_5 = \{0,1,2\} & \\ A_3 = \{0,1\} & & A_6 = \{0,1,2,3\} & \end{array} \quad . \quad (4.1)$$

Suppose the locus of points <u>not</u> in the constrained acceptance region is defined by the explicit edits

$$\begin{array}{l} e_1 = A_1 \times \{0,1\} \times \{0\} \times A_4 \times \{0,1\} \times A_6 \\ e_2 = \{1\} \times A_2 \times \{1\} \times \{0,1\} \times A_5 \times \{2,3\} \\ e_3 = \{0\} \times \{1,2\} \times A_3 \times \{1,2,3\} \times A_5 \times A_6 \\ e_4 = A_1 \times \{0,2\} \times A_3 \times A_4 \times A_5 \times \{0,1\} \\ e_5 = \{1\} \times A_2 \times A_3 \times \{0\} \times \{1,2\} \times A_6 \end{array} \quad (4.2)$$

These are normal-form edits since they are formed by set theoretic Cartesian products. If the edit e_j is formed with A_j, it is said that field i is "not active" in edit e_j. An inactive field's value on a particular record can not cause the failure of the edit in which it is inactive. For example, the record $\underline{y}$ = (0,1,0,0, 1,0) fails only edit e_1 (i.e., $\underline{y} \varepsilon e_1$ in the point set sense). Since fields 1,4 and 6 are not active in edit 1, one or more of fields 2, 3, and 5 in $\underline{y}$ cause the failure.

One set of alternative c_i weights for a given record could be the c_i defined by

$$c_i = 1/(\text{No. times field i active in failed edits} + 1) \; . \quad (4.3)$$

Logically, the denominator of the above expression should be loosely related to the posterior probability of error in field i, at least in the case of "well-behaved" acceptance regions. Then the proposed c_i would be inversely related to this probability, as one desires for the MWFI minimization.

A second set of alternative weights c_i for a given record might be defined by taking

$$c_i = \frac{(\text{No. times field i active in passed edits}+1)}{(\text{No times field i active in failed edits}+1)} \quad (4.4)$$

Here one considers both information which acts to augment the posterior probability of field i error (activity in a failed edit) and information which acts to depress this probability (activity in a passed edit). Intutively this alternative better utilizes our partial information.

There is one very important step that must be taken before the c_i weights defined by either (4.3) or (4.4) can be employed. One can well imagine that the locus of points specified by a set of edits such as (4.2) could easily be specified by several alternative sets of edits. One must insist on the invariance of the weights from (4.3) or (4.4) for different representations of the edits. For normal-form edits, Liepins (1980) has shown that this requires the derivation of a disjoint-sufficient collection of edits, followed by a maximal collection of edits generated as a union of edits from the disjoint-sufficient collection whenever such a union results in a normal-form edit. It can be shown that the explicit edits of (4.2) result in a set of 18 maximal edits.

5. A SIMULATION STUDY

This section reports on the procedure and results in an extensive simulation comparing three sets of alternative c_i weights in the MWFI problem. These alternatives are defined as follows:

Method 1: $c_i = 1$ for all i. This is equivalent to assuming equal prior probabilities of error in each field.

Method 2: c_i defined by (4.3), with a scaling adjustment so that $\min_i c_i = 1$.

Method 3: c_i defined by (4.4), with a scaling adjustment so that $\min_i c_i = 1$.

5.1. Procedure

A flow chart of the procedure executed to simulate the use of the three alternative sets of c_i weights in the MWFI error localization problem is given below. Each execution of the procedure produced 10,000 records that failed one or more of the explicit edits in (4.2) by perturbing (adding errors to) generated records that failed none of the explicit edits. Thus, the fields in error were known deterministically and could be compared against the various MWFI solutions.

The results in this paper were produced by the execution of the procedure three times under the following design:

A. Probabilities of error in each of 6 fields .05 during error generation.

B. Probabilities of error in each of 6 fields .10 during error generation.

c. Probabilities of error in each of 6 fields .20 during error generation.

5.2. Results Format

The simulation results are summarized in Table 1. There is a division of the results based on whether a generated erroneous record failed 2 or fewer maximal edits or failed 3 or more maximal edits. This split was motivated by the pattern one sees - method 1 and method 2 give the same results when 2 or fewer maximal edits are failed. In retrospect, one can demonstrate this fact quite easily conceptually in terms of the miniimization problem as expressed in the abstract.

Before the summary statistics in Table 1 are defined, the following terms must be defined:

Solution = set S of fields to change to minimize objective function subject to changed record being acceptable.

Exact match = solution where S contains all the fields known to be in error, but no other fields.

Partial match = solution where S contains some or all of the fields known to be in error, but no other fields.

Three of the summary statistics in Table 1 are alternative success indices which will range from 0 to 1 - no success to total success to total success, respectively. They have the following definitions:

Average success index 1 = cumulative success index 1 divided by number of records, where cumulative success index 1 is formed by adding one divided by the number of solutions when there is an exact match for a particular record.

Average success index 2 = cumulative success index 2 divided by number of records, where cumulatiave success index 2 is formed by adding the proportion of errored fields identified divided by the number of solutions when there is a partial match for a particular record. Cumulative success index 1 is a subset of cumulative success index 2.

Average success index 3 = cumulative success index 3 divided by number of records, where cumulative success index 3 is formed by determining the proportion of fields correctly dealt with over all solutions, i.e., fields suggested to be in error that were or fields suggested not to be in error that were not. Index 3 is thus a field related index, whereas indices 1 and 2 might be called solution related indices.

Other Table 1 summary statistics are:

Average number matches = cumulative number matches divided by number of records.

Average number solutions = cumulative number solutions divided by number of records.

Matches/solution = cumulative number matches divided by cumulative number solutions.

Four of the above summary statistics, the three success indices and the matches/solution, are provided as conceptually similar statistics for the comparison of the 3 methods of choosing the c_i weights.

5.3 Discussion of Results

The general simulation summary in Table 1 demonstrates the strong dependence of the summary statistics on the probability of error in each field during data generation. All summary statistics deteriorate (i.e., decline, except for average number solutions) noticeably, in the face of increases in this probability except for average number of solutions in the 2 or fewer edits failed case, which is basically constant. It should be noted that the actual proportions of error in the 10,000 generated records were .198, .230, and .295 in the .05, .10, and .20 cases, respectively. Clearly every record must have at least one field in error to fail an explicit edit. Thus, the minimum actual proportion of error in the simulation is 1/6 = .167.

How do the three methods of determining the c_i weights compare? The patterns that arise in the summary statistics of Table 1 may be described in the following ways:

Average number exact matches: Method 1 always produces the most exact matches, followed closely by method 2, while method 3 produces substantially fewer exact matches. This statistic by itself, however, means little in method comparison since the number of solutions produced is a primary factor in finding matches. Method 3 produces relatively more exact matches when 2 or fewer edits are failed, as does method 2 to a lesser extent.

Average number solutions: Method 3 produces many fewer solutions in general, about half the number for each of the other two methods, which do not differ greatly in terms of this statistic. Method 3 produces even fewer solutions in a relative sense when 3 or more edits are failed.

Average success indices: The three success indices follow basically the same pattern. Thus, one may discuss them as a group. Method 3 is always the most successful method as judged by these indices. Method 3's higher success indices come from the records wher 2 or fewer edits are failed. Methods 1 and 2 are practically inseparable based on success indices.

Matches/solution: Method 3 is uniformly judged the best by this statistic regardless of other factors. Method 2 is judged second best by a narrow margin over Method 1.

Overall one sees some statistical evidence that method 3 is an improvement over method 1 while method 2 gives results that look very much like those for method 1. The demonstration of method 3 superiority is hardly conclusive, however, and

certainly not strong enough to indicate great practical significance above and beyond the apparent statistical significance.

6. CONCLUSION

This paper has been a useful piece of research for the authors not so much because of the questions it has answered, but because of the questions it has raised. This is not surprising in retrospect, as the paper reports the first extensive empirical experience with the MWFI error localization process that the authors are aware of.

One question raised centers upon the simulation procedure. Errors were generated for all fields simultaneously in effect. The probability of error in each field was not necessarily as was input to the generation process initially since the generation process was repeated until the record failed one or more explicit edits. Should one drop this stochastic entry of errors which is not really being controlled as one might think in favor of perhaps exhaustive examination of all possible error combinations with all possible acceptable records?

A second obvious question is how meaningful are the success measures that have been utilized? The authors have clearly expressed some uncertainty over this in presenting three separate success indices and the matches/solution statistic.

Many other questions are raised that finally relate to the single broader question, "How specific are the results to the example employed?" There seems to be a need to characterize a particular automatic data editing problem.

In conclusion, it is suggested that there is no question about one issue. An integrated approach to data editing requires that one come to grips with the so-called "error model", both in terms of the characteristics of the error vector $\underline{\varepsilon}$ and the distribution of the true record $\underline{x}$ in the acceptance region.

REFERENCES

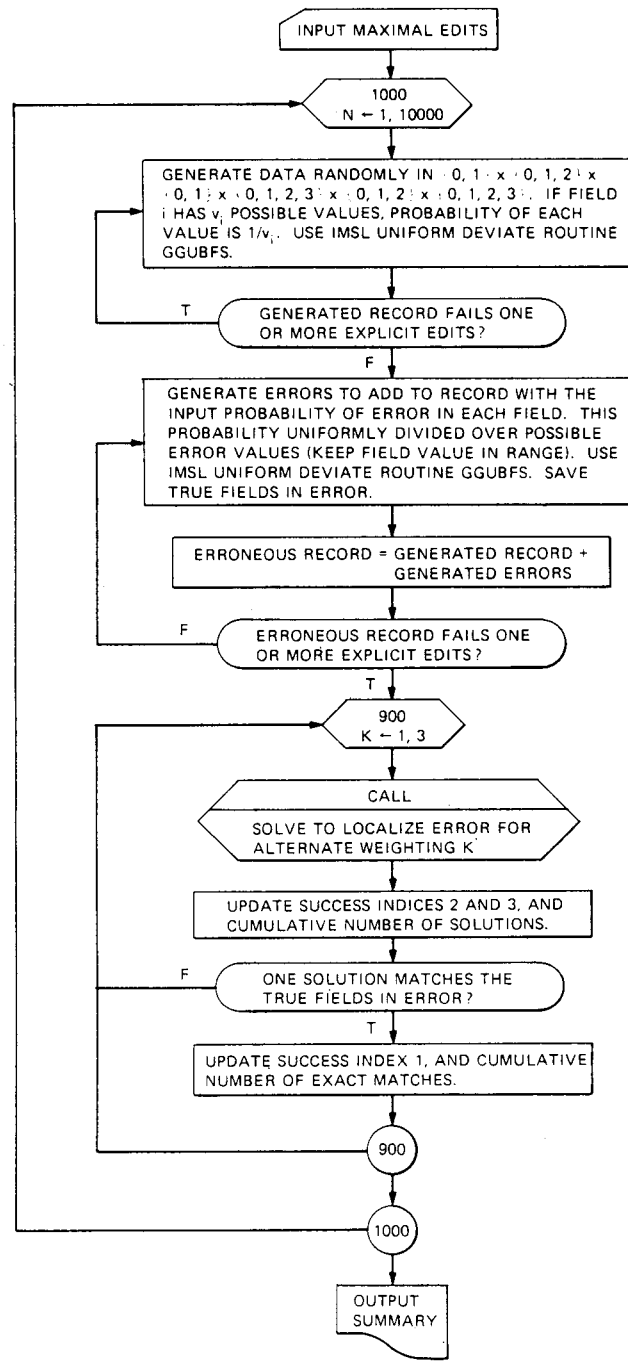
I.P. Fellegi and D. Holt (1976), "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, 71, 17-35.

R. J. Freund and H. O. Hartley (1967), "A Procedure for Automatic Data Editing," Journal of the American Statistical Association, 62, 341-352.

G. E. Liepins (1980), "A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis," Oak Ridge National Laboratory Technical Manuscript, ORNL/TM-7126.

J. I. Naus, T. G. Johnson, and R. Montalvo (1972), "A Probabilistic Model for Identifying Errors in Data Editing," Journal of the American Statistical Association, 67, 943-950.

SIMULATION FLOW CHART



No. Edits Failed	Prob. Error Each Field	Number Records	Method	Average Success Index 1	Average Success Index 2	Average Success Index 3	Average Number Matches	Average Number Solutions	Matches/ Solution
			1	.396	.437	.799	.848	2.36	.360
	.05	8391	2	.396	.437	.799	.848	2.36	.360
			3	.433	.478	.814	.509	1.22	.418
2			1	.333	.409	.776	.712	2.34	.304
or	.10	8160	2	.333	.409	.776	.712	2.34	.304
fewer			3	.363	.444	.789	.430	1.22	.353
			1	.228	.354	.724	.488	2.32	.210
	.20	7664	2	.228	.354	.724	.488	2.32	.210
			3	.249	.383	.737	.292	1.21	.241
			1	.410	.451	.790	.828	2.38	.348
	.05	1609	2	.409	.451	.790	.804	2.21	.364
			3	.412	.453	.791	.421	1.07	.396
3			1	.327	.396	.750	.712	2.71	.263
or	.10	1840	2	.324	.394	.749	.668	2.41	.278
more			3	.331	.396	.748	.343	1.10	.312
			1	.192	.310	.675	.484	3.17	.153
	.20	2336	2	.192	.309	.676	.436	2.71	.161
			3	.206	.326	.680	.223	1.15	.194
			1	.399	.439	.798	.845	2.36	.358
	.05	10000	2	.398	.439	.798	.841	2.33	.360
			3	.430	.474	.810	.495	1.19	.415
			1	.332	.407	.771	.712	2.41	.296
General	.10	10000	2	.331	.406	.771	.704	2.35	.299
			3	.357	.435	.782	.414	1.20	.346
			1	.220	.344	.712	.487	2.52	.193
	.20	10000	2	.219	.343	.713	.476	2.41	.197
			3	.239	.370	.724	.276	1.20	.230

1. General simulation summary.