IMPLIED EDIT GENERATION AND ERROR LOCALIZATION FOR RATIO AND BALANCING EDITS

Maria Garcia[1]
U.S. Bureau of the Census, Washington, D.C., 20233
Maria.M.Garcia@census.gov

Abstract

The U.S. Census Bureau has developed SPEER software that applies the Fellegi-Holt editing method to economic establishment surveys under ratio edit and a limited form of balancing. It is known that more than 99% of economic data only require these basic forms of edits. If implicit edits are available, then Fellegi-Holt methods have the advantage that they determine the minimal number of fields to change (error localize) so that a record satisfies all edits in one pass through the data. In most situations, implicit edits are not generated because the generation requires days-to-months of computation. In some situations when implicit edits are not available Fellegi-Holt systems use pure integer programming methods to solve the error localization problem directly and slowly (1-100 seconds per record). With only a small subset of the needed implicit edits, the current version of SPEER (Draper and Winkler 1997, upwards of 1000 records per second) applies ad hoc heuristics that finds error-localization solutions that are not optimal for as much as five percent of the edit-failing records. To maintain the speed of SPEER and do a better job of error localization, we apply the Fourier-Motzkin method to generate a large subset of the implied edits prior to error localization. In this paper, we describe the theory, computational algorithms, and results from evaluating the feasibility of this approach.

Keywords: editing, error localization, Fellegi-Holt model

1. Introduction

In economic surveys and censuses, survey data files may contain a large number of records with erroneous, missing, or inconsistent data. Errors can arise during data collection due to item non-response, misunderstanding of a survey question or problems with computer data entry. Survey data editing is the process of identifying and correcting errors or inconsistencies in the collected data. Computer records with erroneous or inconsistent data must be edited before the agency produces and publishes relevant and accurate statistics. In statistical agencies, data editing uses a considerable amount of the survey resources available for the publication of statistics. This cost can be reduced if we have an automated system that can be reused by various separate surveys. Currently, for most surveys, the detection and correction of erroneous data is done using an automated software. Fellegi and Holt (Fellegi and Holt, 1976) provided the theory and methodology for the creation of such a system.

An automated system based on the Fellegi-Holt methodology must satisfy the following three requirements (Fellegi and Holt, 1976):

1. The data in each record should be made to satisfy the edits by changing the fewest possible fields.
2. The imputation rules should derive automatically from the edit rules.
3. Imputation should maintain the joint distribution of the variables (fields).

This model requires that the data in each record should be made to satisfy all edits by identifying and changing the minimum possible fields (number one above.) This criterion is referred to as the error localization problem. Fellegi and Holt showed that the implicit edits that can be logically derived from the set of analyst's supplied explicit edits are needed for solving the error localization problem. The complete set of explicit and implicit edits is sufficient to determine imputation intervals for erroneous fields so that an edit failing record is corrected. Prior edit models would fail because they lack the needed information about the original set of explicit edits that may not fail but might fail the imputed record if information in the complete set of edits is not used during error localization.

Several Fellegi-Holt computer systems are currently available for editing continuous economic data: Statistics Canada's Generalized Edit and Imputation System (GEIS) (Cotton, 1999), Statistics Netherlands CherryPi (De Waal, 1996), National Agricultural Statistical Service's AGGIES (Todaro, 1999) and the US Census Bureau's Structured Program for Economic Editing and Referrals (SPEER, Draper and Winkler (1997)). The GEIS, CherryPi and AGGIES software solve simultaneous linear inequality edits using integer programming techniques to implicitly generate the failing implied edits

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

(Rubin, 1975) needed for finding error localization solutions. The SPEER system is used for economic data under balancing and ratio edits and applies simple heuristics to generate a subset of the implicit edits needed for solving the error localization problem. A more detailed description of the SPEER software is given in the next section.

In this paper we applied the Fourier-Motzkin elimination method (Duffin, 1974) to generate a large subset of the implicit edits prior to error localization in the SPEER editing system. In the following sections we present the theory, computational algorithms, and results from using this approach.

2. Implicit Edit Generation and the SPEER edit system

2.1 The SPEER editing software

The Census Bureau has an editing system, SPEER (Structured Programs for Economic Editing and Referrals), for editing continuous economic data that must satisfy ratio edits and a limited form of balancing. The SPEER system has been used at the Census Bureau on several economic surveys since the early 1980's (Greenberg and Surdi, 1984).

This paper describes modifications to the SPEER edit software that maintain the exceptional speed of the system and do a better job of error localization. The current version of SPEER consists of a main edit program and four auxiliary modules. The FORTRAN code for the edit checking, error localization, and imputation routines in the main edit program is new. The four auxiliary modules perform different tasks: the first module automatically determines the bounds for the ratio edits (Thompson and Sigman, 1996); the second module checks the logical consistency of the user supplied explicit edits and generates the implicit ratio edits needed for error localization; the third module generates the regression coefficients that are used in the imputation module; and a new fourth module generates a subset of the implicit linear inequality edits that arise when combining ratio edits and balance equations.

The SPEER software identifies and corrects erroneous fields in data records that must satisfy ratio edits and single level balancing. By single level balancing we mean that data fields (details and totals) are allowed to be restricted by at most one balance equation. It is known that more than 99% of the data items in economic surveys are required to satisfy either no balance equation or single level balancing. A record with n data fields in a computer data file is represented by $v = (v_1, v_2, \ldots, v_n)$. A ratio edit is the requirement that the ratio of two data items is bounded by lower and upper bounds,

$$l_{ij} \le v_i / v_j \le u_{ij},$$

where l_{ij} and u_{ij} are the largest lower bound and smallest upper bound respectively. The bounds can be determined by analysts through use of prior survey data. A balance edit is the requirement that two or more details and a reported total satisfy an additivity condition of the form

$$\sum_{k \in S} v_k - v_t = 0,$$

where S is a proper subset of the first n integers and $t \notin S$. Fellegi-Holt editing model guarantees that if the complete set of explicit and implicit edits is available then we can determine a minimum number of fields to change so that an edit failing record satisfies the edits. In the earliest versions of SPEER which used ratio edits only, it is straightforward to generate the complete set of ratio edits. Since the complete set of explicit and implicit edits is available, it is easy and exceptionally fast to solve the error localization problem.

In the most recent version of SPEER, Draper and Winkler (1997) generate implicit edits induced by failing ratio edits and balance equations "on the fly" for every failing record. The induced edits are then used to further restrict imputation intervals than the restrictions placed by ratio edits only. The solution however, is not necessarily an error localization solution since not all implicit edits are available. This is true in most cases: in general for continuous data it is not possible to generate all the implicit edits for a set of explicit linear inequality edits (Sande, 1978). Recently, Winkler and Chen (2002) provided extensions to the theory and computational aspects of the Fellegi-Holt editing model for discrete data. In their research on discrete data they showed that if most of the implicit edits are computed prior to automatic editing, then error localization algorithms are faster than direct integer programming methods for solving the error localization problem. These results can be extended to continuous numeric data. The main purpose of this paper is to use this idea in SPEER editing when a large subset, but not all, of the implicit edits are generated prior to editing.

2.2 Implicit Edit Generation for Balancing and Ratio Edits

The SPEER edit system has an auxiliary module for generating all the implicit ratio edits for a given set of explicit ratio edits. In the earlier version of SPEER, the needed implicit edits implied by failing ratio edits and a balance equation are generated on the main program for every failing record. This means many implicit edits are repeatedly computed. The new SPEER software generates a large subset of the implied edits prior to SPEER editing. The implied edits are then available to be used in the main edit program and it is not necessary to repeatedly generate the same implicit edits for every edit failing record. This eliminates the need for implicit edit generation during the more computationally intensive error localization program. We want to point that it is feasible to generate implicit edits for SPEER algorithms because it deals with numeric data under ratio edits and single level balancing only. In most situations, implicit edits are not generated because the generation requires days-to-months of computation. For example, for the Italian Labor Force Survey (Barcaroli and Ventura, 1997) the amount of computation would be prohibitive, with estimates of at least 800 days on the largest IBM mainframe for the edit generating algorithms (Winkler, 1997).

The new added module for generating implicit linear inequality edits for ratio edits and balancing edits is based on the Fourier-Motzkin elimination method (Duffin, 1974). This methodology has been used in new algorithms for the Leo editing system developed at Statistics Netherlands (Quere, 2000). The Leo software uses Fourier-Motzkin elimination to delete a field from nodes representing the current set of edits in a tree search algorithm for solving the error localization problem.

The mathematical knowledge to develop and understand the implicit edit generation is simple. The method developed by Fourier for checking the consistency of a set of inequalities can be used to generate implicit linear inequality edits. Suppose we have a ratio edit $l_{ij} \le v_i / v_j \le u_{ij}$ and balance equation $\sum_{k \in S} v_k - v_t = 0.$ Using simple algebra we can rewrite the ratio edit as two linear inequality edits and the balance equation as two linear inequality edits. If we can find a variable in common in the linear inequality edits corresponding to the ratio and balance edits, say $k = i$ for some $k \in S,$ and provided the coefficients of the common variable have opposite signs, then we can eliminate the common variable by creating a linear combination of the two edits. For example, if $-v_1 + l_{14}v_4 \le 0$ and $v_1 + v_2 - v_3 \le 0$ are linear inequality edits derived from the ratio and balance equation respectively, then $v_2 - v_3 + l_{14}v_4 \le 0$ is a new implied edit. The new SPEER implicit edit generation algorithm uses this methodology to generate as many implicit edits as possible from linear combinations of the complete set of ratio edits and the balance equations. The algorithm is repeated to generate new implied edits from linear combinations of the newly generated implicit edits and the current set of edits. Generating a large subset of the implicit edits using this methodology has numerous advantages. For ratio edits and single level balancing the edit generation is fast, the logic is simple, and the availability of implicit edits prior to editing reduces computational effort during error localization. The reduction can be so significant that the speed of the main edit program is no longer an issue when compare to Chernikova-type error localization algorithms. This is very important since reducing computations is a critical aspect of designing a Fellegi-Holt system.

While doing this research we found that it is possible that the ratio edits bounds in the complete set of ratio edits are not necessarily optimal. This could happen when there are two details required to balance to a reported total and two terms of this balance equation are in the ratio edit. Consider the following example, the coefficients of v_2 have opposite signs in edits $v_1 - u_{12}v_2 \le 0$ and $v_1 + v_2 - v_3 \le 0$. Using Fourier elimination we can generate implicit edit $v_1 - \frac{u_{12}}{1+u_{12}} v_3 \le 0.$ In this case, if $\frac{u_{12}}{1+u_{12}} \le u_{13}$, then the upper bound u_{13}, for the ratio edit connecting fields v_1 and v_3 is not optimal and needs to be adjusted.

We note that since any pair of ratio edits with a common data field implies another ratio edit, updating at least one bound in the complete set of edits implies that all lower and upper ratio edit bounds must be revised and updated. In a very simple example with four fields and six ratio edits in the complete set of ratio edits, we found that six of the twelve lower and upper bounds needed to be changed after two passes through the new implicit edit generation program. The possibility that the ratio edits bounds should be modified using the edit restrictions imposed on data items by the balance equations have not been considered in the earlier version of the SPEER edit system. It implies that the algorithms in the previous version of SPEER did not have available the edits that impose the most restrictions on the data fields, and therefore could change the error localization solutions and the imputation intervals used to "fill-in" data in the imputation algorithms.

The implicit edits generated by ratio edits and balance equations are computed using the methodology described above. The code is written in SAS and SAS/IML. The input of the new implicit edit generation module is the complete set of ratio edits and the balance equations. The algorithm used in the implicit edit generation is as follows:

Step 1. Represent the ratio edits and balance equations as homogeneous linear inequality edits, $Av \le 0,$ $A = \begin{pmatrix} R \\ B \end{pmatrix},$ where R and B are the matrices of coefficients corresponding to ratio and balance edits respectively, and v is the vector of data fields.

Step 2: Choose two linear inequality edits with a common data field v_k in which the coefficients of v_k have opposite signs. Use Fourier-Motzkin elimination to generate a new implied edit.

Step 3: Verify that the new implied edit is an essentially new derived edit. If the new implied edit has only two entering fields then check whether the corresponding ratio edit bound needs to be updated. If any ratio edit bound is updated then revise and update the complete set of ratio edits.

Step 4: Adjoin the coefficients from the new implied edits to the matrix of coefficients A, and go to Step 2.

2.3 Editing in the new SPEER

The current version of SPEER (Draper and Winkler, 1997) for editing numeric data under ratio edits and single level balancing generates failing implicit edits during error localization for every edit failing record. In the previous section we described how the Fourier-Motzkin elimination method can be used to calculate linear inequality edits implied by ratio and single-level balancing edits. In the new version of SPEER we use this methodology to generate a large subset of the implicit edits prior to automatic editing which considerably simplifies error localization in the SPEER edit system. This is important because the implicit edits are then available to be used many times in the error localization routine for every edit failing record. The need to repeatedly generate the implicit edits for every edit failing record is eliminated and the computational effort during error localization is reduced.

In the new version of SPEER, the edit checking, the error localization, and the imputation modules have all been rewritten to use the implicit edits generated prior to automatic editing. The edit checking routine identifies the records failing any ratio edit, balance equation, or implicit edit. Changes to the edit checking routine are straightforward, we simply added code to determine if any of the implicit edits generated using the new algorithm failed. The code in the previous version of the error localization module needed to generate and error localize failing implied edits was not particularly easy, and it is no longer needed. Error localization has been greatly simplified. For every data record marked as failing at least one edit (ratio, balance or implicit) in the edit checking routine, the error localization module uses a greedy algorithm (Nemhauser and Wolsey 1987) to determine the minimum number of fields to impute so that the record no longer fails.

The code in the imputation algorithm also uses the information from the implicit edits generated prior to automated editing. We recall that one of the main results of the Fellegi-Holt (Fellegi and Holt, 1976) theory is that if we know the values of a subset of fields that satisfy all edits that place restrictions on those fields only, then we can impute for the remaining fields so that the record satisfies all edits. The imputation routine will successively check each field identified to be changed and impute for that item. As in the previous version of SPEER, if there is only one term in a balance equation marked for imputation, then the balance equation is used to impute the value of the item. Otherwise, we impute a field value using the information from the other known fields' values, the ratio edits restrictions, balance edits and implied edits to determine the interval into which to impute.

The algorithm is as follows:
For each data record do,

Step 1: Use ratio edits, balance equations, and implicit edits generated using the implicit edit generation algorithm in Section 2.2 to identify failing edits. If record fails at least one edit, continue. Otherwise, go to the next record.

Step 2: If the record fails at least one edit, use the failing ratio edits, failing balance equations, and the failing implicit edits identified in Step 1 in a greedy algorithm to determine a minimum number of fields to be changed so that the record satisfies the edits.

Step 3: For each field marked to be imputed in Step 2, use the other known fields (reported and imputed) and the edit restrictions to determine an interval into which field values can be imputed.

Section 3: Preliminary Results

Our initial test runs used two one-industry test data sets: the first data set consists of six ratio edits and one balance equation in four fields (Winkler and Draper, 1996); the second test data set is a one-industry subset of the 1997 Annual Survey of Manufactures (ASM) with 136 ratio edits and two balance equations in 17 fields. We first run the implicit edit generation program. The set of implicit linear inequality edits generated using this program is then used, along with the complete set of ratio edits, as input to the new SPEER system. In the previous section we mentioned that it was possible that the ratio edit bounds need to be adjusted during implicit edit generation –this is important since the ratio edits bounds are used for computing imputation intervals so that record no longer fails. Table 1 displays the total number of ratio edit bounds adjusted after one and two passes through the implicit edit generation program. For the first edit set, six (out of 12) ratio edit bounds were adjusted while for the ASM edits a total of 52 ratio edit bounds (out of 272) were adjusted after two passes through the implicit edit generation program.

Table 1: Number of Adjusted Ratio Edit Bounds

Data Set	Number of Items	Number of Ratio Edits	Number of Bounds Adjusted After One Pass	Number of Bounds Adjusted After Two Passes
Winkler, Draper (1996)	4	6	6	6
1997 ASM	17	136	12	52

For the first test data (four items), the check edit module identified five records with either erroneous or missing data. For each record, the imputation routine determined imputation intervals and successfully imputed all field values identified to be changed during error localization. We used the test data from the 1997 ASM (17 items) for comparing the results when running the 1997 version of SPEER (SPEER' 97) and the new version of SPEER (SPEER' 02). Both programs identified the same 76 edit failing records, however the results of the error localization routines are different. Table 2 displays the items identified to be imputed and the number of times each item was deleted (marked to be changed). The total number of times a field was marked for deletion during error localization is consistently higher –except for one item (WW), in SPEER' 97 when compare with SPEER '02. This result is expected. In Section 2 we mentioned that SPEER' 97 does not necessarily error localize since not all implicit edits are available. SPEER' 02 has more information available from the set of implicit edits generated prior to error localization, therefore it should do a better job of error localization and preserve more of the reported data.

Table 2: Number of Times Field was Deleted for One Industry in 1997 ASM Data

ASM fields	SPEER' 97	SPEER' 02
SW	26	24
TE	16	3
WW	4	5
OW	25	22
PW	6	5
OE	21	19
LE	3	2
VP	7	3

–Only fields where there was a difference in the number of times field was marked for deletion are listed.

4. Discussion

The purpose of this research was to develop a new implicit edit generation algorithm for the SPEER edit system based on the Fourier-Motzkin methodology for finding solutions to a system of linear inequality edits. The system takes as input the complete set of ratio edits and the balance equations. The set of ratio edits and balance equations are then represented as linear inequality edits. These linear inequality edits are then used to generate implicit edits. We recall that the newly generated implicit edits can be combined with the initial set of edits to generate a larger subset of implicit edits. The implicit edits that are generated need to be checked and any redundant edits are discarded. The software has an option for choosing the maximum number of passes through the system.

In the previous section we presented preliminary results from testing the algorithms described in this paper. Using this methodology has several potential advantages for Census Bureau's SPEER editing system. First, the logic needed to implement the algorithm for the edit generation system and SPEER editing are simple, easy to understand and can be used with any survey under ratio edits and single level balancing. Another advantage of using this new algorithm is that the implicit edits generated prior to SPEER editing are available to be used repeatedly during error localization. This greatly reduces the computational effort in the error localization module since there is no need to compute failing implicit edits for every edit failing record. This is a particular strength. This approach is not however without its disadvantage: generating a large subset of implicit edits for some surveys could possibly need large data structures since the set of implicit edits can grow very large.

The results from these initial test runs are very encouraging. The modifications proposed for the SPEER editing system are still in the testing phase. For now, the recommendation is to do more testing. The benefits of using this approach are now being tested with the 1997 Annual Survey of Manufactures full production edits and data.

4. References

Barcaroli, G., and Venturi, M., (1993), "An Integrated System for Edit and Imputation of Data: an Application to the Italian Labor Force Survey," *Proceedings of the 49th Session of the International Statistical Institute*, Florence, Italy, September 1993.

Cotton, C. (1999), "Functional Description of the Generalized Edit and Imputation System," Statistics Canada Technical Report.

DeWaal, T., (1996), "CherryPi: A computer program for automatic edit and imputation," *UN work Session on*

Statistical Data Editing, November 1996, Voorburg.

Draper, L., and Winkler, W., (1997), "Balancing and Ratio Editing with the New SPEER System," *American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods*, 582-587.

Duffin, R. J., (1974), "On Fourier's Analysis of Linear Inequality Systems," *Mathematical Programming Study*, North-Holland Publishing Company, 71-93.

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

Greenberg, B. and Surdi, R., (1984), "A Flexible and Interactive Edit and Imputation System for Ratio Edits," SRD report RR-84/18, US Bureau of the Census, Washington, D.C.

Nemhauser, G. L. and Wolsey, L. A., (1988), "*Integer and Combinatorial Optimization*,"John Wiley: NY.

Quere, R., (2000), "Automatic Editing of Numerical Data," *Report, Statistics Netherlands*, Voorburg.

Rubin, D. S., (1975), "Vertex Generation in cardinality Constrained Linear Programs," *Operations Research*, No. 23, 555-565.

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

Thompson, K. J. and Sigman, R. (1996), "Statistical Methods for Developing Ratio Edits Tolerances for Economic Censuses," *American Statistical Association, Proceedings of the Section on Survey Research Methods*.

Todaro, T. (1999). "Evaluation of the AGGIES Automated Edit and Imputation System," National Agricultural Statistics Service, USDA, Washington D. C., RD Research Report No. RD-99-01.

Winkler, W., (1997) "Set Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods.*

Winkler, W. and Chen, B. C. (2002), "Extending the Fellegi-Holt Model of Statistical Data Editing," *SRD research report Statistics* RRS2002/02, US Census Bureau, Washington D.C.

Winkler, W. and Draper, L. (1996), "Application of the SPEER Edit System," *SRD research report*, RR96/02, US Census Bureau, Washington D.C.