EDITING ECONOMIC DATA

John Kovar, Statistics Canada and William E. Winkler, Bureau of the Census
William E. Winkler, Bureau of the Census, Washington, DC 20233-9100, bwinkler@census.gov

KEYWORDS: Fellegi-Holt model, imputation, nearest neighbor

Two computer systems are currently available for editing continuous, economic data: Statistics Canada's General Edit and Imputation System (GEIS) and the Census Bureau's Structured Programs for Economic Editing and Referrals (SPEER). GEIS, the more general of the two systems, uses linear inequality edits and provides several imputation options. SPEER uses ratio edits which are a special case of linear inequality edits and provides only one imputation option. The main algorithm of GEIS is a relatively fast extension of Chernikova's algorithm for finding extreme points in n-dimensional space. SPEER edits are exceedingly fast because they do not involve advanced integer programming algorithms. The methods of the two systems are described and an empirical comparison is provided.

1. INTRODUCTION

Computer files used for administrative or survey purposes may contain large numbers of records, some of which contain logical inconsistencies or incorrect data. Pritzker, Ogus, and Hansen (1965) describe the nature of the problem. Errors can arise because methods of creating records in files are not consistent, because questions are not understood, or because of transcription or coding problems. In many situations, data files are edited using custom software that incorporate rules developed by subject-matter specialists. If the specialists are unable to develop the full logic needed for the edit rules, then the subsequent edit software is in error. If programmers do not properly code the rules, then the software would be in error. Developing software from scratch each time a data base is redesigned is time-consuming and error-prone. It is better to have a system that can describe edit rules in tables that are read and utilized by reusable software modules. The tables could be more easily updated and maintained than complex if-then-else rules in computer code. The software would automatically check the logical validity of the entire system prior to the receipt of data during production processing.

Fellegi and Holt (1976), hereafter FH, provided the theoretical basis of such a system. FH had three goals that we paraphrase:

1. The data in each record should be made to satisfy all edits by changing the fewest possible variables(fields). (FH1)
2. Imputation rules should derive automatically from edit rules. (FH2)
3. When imputation is necessary, it should maintain the joint distribution of variables. (FH3)

The key to the FH approach is understanding the underpinnings of goal 1. Goal 1 is referred to as the *error localization* problem. In the FH model, a subset of the edits that can be logically derived from the explicitly defined edits (called *implied or implicit edits*) are sufficient to solve the error localization problem.

The purpose of this paper is to describe two automatic edit and imputation systems that have been developed for continuous, economic data and to provide some empirical results. The first system is the Census Bureau's Structured Programs for Economic Editing and Referrals (SPEER) that performs ratio editing and balancing (assuring that items add to totals). The second system is Statistics Canada's Generalized Edit and Imputation System (GEIS) based on linear inequality edits. GEIS is more general because linear inequality edits include ratio edits and balance edits.

The outline of the paper is as follows. In section two, we give some background on the Fellegi-Holt model of editing. The third and fourth sections describe SPEER and GEIS, respectively. In the fifth section, we describe the empirical data that are used in comparisons and, in the sixth section, we provide the results of these comparisons. The final two sections consist of a discussion and a summary.

2. THE FELLEGI-HOLT MODEL OF EDITING

Fellegi and Holt (1976) were the first statisticians to show that implicit edits are sufficient to determine ranges for imputed values that would satisfy the edits. Prior edit models failed because they only made use of explicit edits. The implicit edits contain essential information about explicit edits that may not fail for a record that fails other explicit edits and are needed for imputing values so that the original non-failing edits do not fail the record with the newly imputed values. We

denote the set of explicit edits plus the set of implicit edits needed for error localization by E^c. Let Ω_K be the subset of E^c that involves only fields 1, 2, ..., K. The following theorem is the main error localization result of FH.

Theorem 1 (FH). If y_i^o, I = 1, 2, ..., K-1, are, respectively, some possible values of the first K-1 fields, and if these values satisfy all edits in Ω_{K-1}, then there exits some value y_K^o such that y_i^o, I = 1, 2, ..., K, satisfies all edits in Ω_K.

By reasoning inductively, we can fill in y_i^o, I = 1, 2, ..., K-1, with values y_i^o, I = K, ..., N, such that y_i^o, I = 1, ..., N, satisfies all edits in E^c. Since the ordering is arbitrary, we can assume that any subset s and any set of values y_j^o, $j \in s$, that satisfy edits in E^c with entering fields in s can be completed to a record that satisfies all edits in E^c. In the imputation terminology of Little and Rubin (1987), a completed record is one in which all missing items are filled in. In edit/imputation, we fill in items if they are blank and can fill nonblank items with replacement value-states if the original nonblank items are involved with edits that fail. If $r = \{y_i^o, I = 1, ..., N\}$ is a record that fails a set of edits E and s is the set of fields that enter the edits in E, then we can find a minimal cardinality subset s_1 of s so that $\{y_j^o, j \notin s_1\}$ can be completed to a record that satisfies all edits. If we consider weights c_i, I = 1, ..., N, then we can find the minimal weighted subset s_1 of s. We observe that E^c is a set of edits that is sufficient for determining the minimal number of fields (i.e., the set s_1) that must be imputed to change (complete) an edit-failing record to one that satisfies all edits.

In practice, generating implied ratio edits is easy and different groups have been able to develop FH ratio edit systems. While generating implicit linear inequality edits is straightforward, practical error localization is not easy because of the large number of implicit linear inequality edits and the exorbitant amount of computation needed for an integer-programming solution to the error localization problem.

3. SPEER

The SPEER edit system is designed for ratio edits of continuous economic data. The first version of SPEER was written by Brian Greenberg (Greenberg and Surdi 1984, Greenberg and Petkunas 1990) and the current version was written by William Winkler (1996). The computational algorithms, much of the imputation methodology, and the FORTRAN source code in the current version are new. If variables are defined by V_i, I = 1, ..., N, then ratio edits take the form:

$$L_{ij} < V_i / V_j < U_{ij} \tag{3.1}$$

and balance edits take the form

$$\sum_{I \in S} V_i - V_j = 0, \tag{3.2}$$

where S is a proper subset of the first N integers and $j \notin S$. Simple algebra allows the reexpression of the two ratio inequalities in (3.1) as two linear inequality edits and the equality in (3.2) as two linear inequality edits. The bounds L_{ij} and U_{ij} are determined by analysts through use of prior data. A special methodology and program D_MASO facilitates analysts' determination of bounds. A newer bound-determination methodology based on the Exploratory Data Analysis technique of bounded fences (Thompson and Sigman 1996) appears to give somewhat better bounds than D_MASO and requires less human intervention.

The current version of SPEER only allows individual fields to be restrained by at most one balance equation. Extensive review of the edits in use for economic surveys at the Census Bureau has shown that well over 99% of fields in different surveys need to be restrained by one or fewer balance equations. Whereas creating algorithms and writing software for general, multilevel balancing are quite difficult, the algorithms and computer code associated with the one level of balancing in SPEER are quite straightforward.

SPEER software consists of three main programs. The first generates implicit edits (bounds) and checks the logical consistency of the ratio edits only. An auxiliary simplex program checks the logical consistency of the set of ratio and balance edits. The second program generates regression coefficients for the equation $V_1 = ß_{12} V_2 + \varepsilon$ that are used in the imputation module of the main SPEER program. The main SPEER program also uses the implicit edits and the raw data file as inputs. Prior to imputation, the main SPEER program generates failed implicit edits that can be derived from combinations of ratio and balance edits. These extra implicit edits, which we call *induced edits*, are used to restrict imputation ranges further than the restrictions placed by ratio edits only. The induced edits assure that imputed values satisfy ratio and balance edits.

Due to the simplicity of algorithms, SPEER code is exceedingly fast. Generating 272 pairs of implicit edit bounds in each of 546 industrial categories requires a total of 35 seconds on a Sparcstation 20 and 115 seconds on a 75 MHZ Pentium. Because ratio edits are inherently straightforward, most SPEER code is easy to understand and maintain. The code is completely

portable. Using SPEER on other machines merely requires copying FORTRAN source code and recompiling it. Documentation is minimal, consisting primarily of instructions on how to run the code. SPEER documentation and source code are free and available from the second author by request.

4. GEIS

Statistics Canada's Generalized Edit and Imputation System (GEIS) adheres to the FH tenets as does the SPEER system. The solutions, however, are quite different. The first prototype of the system was created by Sande (1979)~ the current version of GEIS is documented in Kovar, MacMillan and Whitridge (1991).

First, subscribing to the FH assumption that errors happen at random and with relatively low probabilities, we conclude that joint probability of multiple errors is very low. This makes the first FH tenet attractive: to minimize the number of fields to impute, or, optionally a weighted number of fields. We note, however, that GEIS solves this problem without explicitly generating all the implied edits, unlike SPEER.

Secondly, in GEIS, the imputation problem is addressed using nearest neighbor (hot deck) method. (Other imputation methods are also offered for the convenience of users.) While the nearest neighbor method does not explicitly make use of the edits to generate the imputations, the edits are used to identify the clean records which can be used for imputation, thus satisfying the second FH tenet, that is, that, the subject matter officers need not explicitly generate the (if-then-else) imputation rules.

Thirdly, as all variables that need imputation for a given record are (usually) taken from the same donor record, and since the plausibility of the imputed values is verified by checking that the would-be imputed record passes all edits, we assume that the joint marginal distributions are not perturbed too seriously. As such, the third FH principle is satisfied.

In more detail, first with respect to error localization, GEIS proceeds as follows. The user defines an acceptance region by means of linear inequality edits which establish the relationship among the variables. Linear programming techniques are brought in to establish internal consistency, non-degeneracy, and non-redundancy of the edit set (Kovar, MacMillan and Whitridge 1991). Any n-tuple, corresponding to a given observation, can then be verified to either lie within the acceptance (feasibility) region or not, corresponding to the notion of the record passing the set of edits, or not. For records that do not pass the edits, a minimum number of fields to be imputed (FH1) is identified by means of a modified Chernikova's algorithm (Chernikova 1964, 1965, Sande 1979~ Schiopu-Kratina and Kovar 1989).

Briefly, the edit set can be described as

$$\begin{matrix} AV \le b \\ V \ge 0 \end{matrix} \quad (4.1)$$

where A is a matrix of coefficients of the linear inequalities corresponding to the edits, b is a column vector of the corresponding constants, and V is a column vector corresponding to a given record. For records V which pass all the edits, the system (4.1) is satisfied, for failing records, (4.1) is not satisfied. The problem is to minimize the cardinality (number of nonzero entries) of the correction vectors y and z, such that the dot product of y and z is zero, and,

$$\begin{matrix} A(y-z) \le b - AV \\ y - z \ge -V \\ y \ge 0 \\ z \ge 0 \end{matrix} \quad (4.2)$$

which can be written as

$$A_1 \begin{pmatrix} y \\ z \end{pmatrix} \le b_1 \; ; \qquad \begin{pmatrix} y \\ z \end{pmatrix} \ge 0$$

where

$$A_1 = \begin{pmatrix} A & -A \\ -I & I \end{pmatrix} \qquad b_1 = \begin{pmatrix} b-AV \\ V \end{pmatrix}$$

Sande (1979) and Schiopu-Kratina and Kovar (1989) show that a solution can be obtained using the Chernikova's algorithm, by examining (in a suitably controlled manner) the extreme points of the system (4.2) in the variables (y,z)', since V is a constant referring to the particular failed record in question. Note that missing values can be represented by any value falling outside of the feasible region of system (4.1), for example, by -1. The reader is referred to Schiopu-Kratina and Kovar (1989) for details relating to the actual implementation used in GEIS.

We note again, that GEIS does not make use of implied edits to find the solution to the error localization problem. While the system can generate implied edits (using Chernikova's algorithm applied to the dual system of (4.1)), this is done for diagnostic purposes only. Implied edits are used in GEIS only to let the user get a feel for the edits that were actually specified. In most real applications, the number of

implied edits is much too large to be useful for diagnostics, let alone error localization. Along the same lines, the extreme points of the system (4.1) are generated in order for the user to be able to assess the efficiency of the edits: the extreme points indicate to the user the most extreme records that would pass the edits, and may indicate where more restrictive edits are needed.

Secondly, with respect to imputation, GEIS offers two broad categories of imputation. In either case, the user is not required to write if-then-else imputation rules (FH2). As part of the first category, several "imputation estimators" (GEIS Development Team, 1990) are offered. These allow the user to use historical imputation, ratio imputation, and mean imputation among others, in order to impute the fields identified in the error localization step, one at a time. Note that while often intuitively appealing, these methods do not ensure that the resulting record will pass all the edits. The reader is referred to Kovar and Whitridge (1990) and Cotton (1991) for more details.

The principal method of imputation in GEIS is a hot deck approach using the nearest neighbor methodology. For every record in need of imputation, all clean records are searched, and the closest one to the recipient is used to impute the error localized cells, provided the resulting imputed record passes all edits. If not, the next closest record is tried, and so on. Of particular note is the fact that all variables are donated simultaneously, thus preserving (in principle) the data distributions as much as possible (FH3). As well, since all edits must pass before the imputation is accepted, the resulting records are guaranteed to be 'clean'. (If this is not possible, the record remains unimputed.) Note again, that this is not the case when imputation estimators are used.

In theory, all recipient records must be compared, in terms of a distance, to all potential donors. To reduce this problem to a manageable size, an efficient k-d tree algorithm is used to search the donor population (Kovar and Whitridge 1990, Cotton 1991). The overhead incurred in constructing the search tree is recuperated with even the smallest of donor decks and very few recipient records. Because the k-d tree is constructed by splitting the donor population around the median of the variable with the largest range, an appropriate transformation of the data and a suitable distance measure must be used. For this reason, all of the data are transformed to uniform marginals (standardized rank-order statistics), and the L-infinity norm (also known as the minimax distance) is used (Sande 1979). More precisely, the distance between two records, $V_s = (V_{s1}, ..., V_{sn})$ and $V_r = (V_{r1}, ..., V_{rn})$ is given by

$$d_{sr} = \max_i |t_{si} - t_{ri}| = \lim_{h \to \infty} \sqrt[h]{\sum_{i=1}^{n} (t_{si} - t_{ri})^h} \qquad (4.3)$$

where $t_{si} = F_i^{-1}(V_{si})$ is the transformed value of the i'th variable of the data record V_s in question, and F_i is the empirical distribution function of the i'th variable based on all the useable (non-missing and valid) observations.

The set of fields used to calculate the distance d between a failing record and a clean record is a subset of the fields not identified to be imputed on the recipient record. These fields are referred to as the matching fields in GEIS. The actual set of matching fields is found by means of linear programming techniques which identify the reduced set of edits which are involved in the definition of the acceptance region for the particular failed edit record (with the fields to impute as the only unknowns). Only the active variables not identified for imputation are retained as matching fields. See Schiopu-Kratina and Kovar (1989) for more details.

Finally, we point out that GEIS relies heavily on the data management functions of the ORACLE RDBMS. Currently it runs under MVS, Unix and DOS operating systems, though the applications that would make use of a DOS environment would have to be limited in size. By contrast, applications the size of the Canadian Census of Agriculture (about 300,000 records and in excess of 400 variables) have been successfully processed by GEIS under MVS. To process such a large problem, the Agriculture Census was broken into subcomponents that were processed separately and the subcomponent solutions were combined to get the final results. A confirming run at the end assured that the final results satisfied edits and balance equations. Preprocessing work was done to assure that solutions (possibly suboptimal) could be obtained by the approach of breaking into subcomponents and then recombining. GEIS (without the proprietary source code) is available for 25,000 Canadian dollars, on an institutional licence basis.

5. EMPIRICAL EVALUATION

A portion of a Canadian agriculture survey comprising some 1700 records and 10 variables was segregated from the final, pre-publication files. The variables on the file include a record identifier, the value of land under cultivation (a frame variable assumed known for

the whole population), and the response variables: income, expenses, assets, as well as some of their components (inc_p1, inc_p2, inc_p3, exp_p1 and exp_p2).

Due to confidentiality reasons, this data set could not have been used for this study. Instead, an artificial population resembling the real one was created, by generating the land variable so that its distribution resembles the observed one, matching in particular the mean and the variance. The other key variables were generated conditionally on the land variable, following the relationships observed in the true population. Means, variances, ranges and correlations of the synthetic population resemble those of the true population quite faithfully. One thousand records were created, which satisfy all of the edits specified below. The data set containing this synthetic population is referred to as the 'clean data set'.

Starting with the clean data set, nonresponse and errors were introduced in 30% of the records, with a probability inversely proportional to the value of the land variable. More specifically, the probability of nonresponse or error for the i^{th} record was set to $P_i = 1 - \exp(cx_i)$, where x_i is the value of the land variable for the record in question, and c is a constant calibrated so that an expected proportion of 30% flagged units be attained. This scheme corresponds more closely to reality than a purely random selection, as larger units generally receive more attention during follow-up, and thus tend to contain less errors. For the records that were flagged to be perturbed, one of a number of actions was taken. These actions included deleting one of income, expenses, or assets, or any combination of them, including the possibility of 'total nonresponse', i.e., only the identifier and land variables remaining on the file. About half of the 30% of the identified units were subjected to such incidences of nonresponse. For the remaining 15% of units, errors of various types were introduced. These included switching of variables, destroying the additivity and subadditivity relationships, etc. The land variable was not modified on any record, as it was assumed to be a 'frame' value. Every attempt has been made to ensure that the errors generated resemble those encountered during the actual production, both in terms of quantity as well as type.

The resulting data set was dubbed the 'unimputed file'. As expected, the means of the clean records on the unimputed file are significantly higher than the corresponding means on the clean data set, ranging from an increase of 9% for assets, 12% for expenses, to 13% for income, due to the nonrandom nature of the error generating mechanism. The effect of the errors introduced can be seen in Table 1 below, by comparing columns 1 and 3. Clearly a number of outliers were generated as a result of variable switching - not an unusual situation in practice.

Finally, the following edits were postulated.

1) All fields ≥ 0
2) income = inc_p1 + inc_p2 + inc_p3
3) exp_p1 $\leq$ expenses / 2
4) exp_p2 $\leq$ expenses / 2
5) expenses $\leq$ 1.25 income
6) exp_p1 + exp_p2 $\leq$ expenses
7) (inc_p1 + inc_p2) / income ≥ 0.5
8) 5,000 $\leq$ income $\leq$ 300,000
9) assets $\leq$ 1,000,000
10) land ≥ 10
11) land $\leq$ 1,500
12) income $\leq$ 500 + 3 expense
13) income $\leq$ 500 land

These edits bear a close resemblance to those actually used in production. Some, as edit 6, for example, are redundant, but are specified for ease of readability, rather than mathematical completeness. Others are implied, such as inc_p3 $\leq$ income / 2. This edit set was used within GEIS without modification.

In the case of SPEER, slight modifications were needed because the software only allows for ratio edits and simple types of balancing. In particular, edit 6 was deleted, edit 7 was handled by creating a dummy variable which is a sum of inc_p1 and inc_p2, and, edit 12 was modified to income $\leq$ 3.1 expenses, i.e., income / expenses $\leq$ 3.1. Bounds were handled ahead of time, by setting out of bounds variables to missing.

Both systems generated the fields to impute automatically as described in the preceding sections. Default imputation actions were performed. That is, donor imputation was used in GEIS, and ratio (regression) imputation was performed in SPEER. The resulting data sets are referred to as the 'GEIS imputed' and the 'SPEER imputed' files, respectively.

6. RESULTS

In both cases, the setup of the programs, including the necessary edit modifications, the importing and exporting of the data sets into appropriate formats, variable definition, etc., was completed in less than one day. The actual error localization and imputation runs were completed quickly. SPEER needed approximately 3.8 seconds on a Sparcstation 20 and GEIS needed less than 2 minutes on a Hewlett-Packard G60 which is 20-40% faster than the Sparcstation 20.

In comparing SPEER and GEIS, we considered the number of fields that needed to be imputed and the quality of the aggregate statistics from the final imputed data base. For the number of fields, GEIS performed better. Both GEIS and SPEER identified the same 277

records as failing edits. GEIS and SPEER produced identical sets of fields to impute in 121 cases and equivalent sets of fields to impute in 112 cases. In the remaining 44 cases, GEIS required imputation of one less field than SPEER. All 44 cases were associated with the balance equation INCOME = INC_P1 + INC_P2 + INC_P3.

The imputed files were compared to the clean file in terms of 1) mean differences (i.e., differences of means) in order to verify whether the imputation actions were able to re-establish the true means, 2) correlation structures in order to check whether correlations between key variables were affected, and 3) mean absolute differences (MAD) in order to quantify the performance at the record level. These are summarized in the two tables at the end of the paper.

7. DISCUSSION

With both SPEER and GEIS, the means of the imputed data were quite close to the means of the original clean data. The correlation structure was also preserved, particularly with SPEER whose imputation method takes the correlations into account directly. Results of an experiment (not shown in this paper) in which the nonresponse and errors were generated with equal probabilities (rather than proportionally to the inverse of the land value) were even better. Furthermore, results of a study conducted at Statistics Canada using the real population, and following exactly the same approach as above except that the SPEER runs were not done, were virtually identical, and in most instances better. This indicates that the artificiality of the data set used does not compromise the results of this study in any way.
Both software systems were easy to handle - neither can be preferred on that basis alone. SPEER does not require any additional software and is marginally more portable. SPEER was designed to be part of a general economic edit system that includes many general options for industrial coding and imputation that are specific to Census Bureau surveys. It was designed to have code that can be modified by Census Bureau programmers. SPEER was never designed and documented in a fashion that would make it easily useable by third party users. Due to its limitation to ratio edits, extra work is needed to use SPEER on the empirical example of this paper. Extensive review of more than 100 surveys and censuses in the economic area of the Census Bureau have not identified any situations in which general linear inequality edits are needed. All of these survey systems require ratio and balance edits only.

GEIS has many additional imputation options that SPEER does not have. The use on nearest-neighbor imputation in GEIS allows for easier handling of secondary variables for which explicit ratio relationships, needed by SPEER, are harder to come by and may be better at preserving distributional properties of data in some situations.

The set up time with either SPEER or GEIS is relatively negligible. Both 'imputers' used the software in its most generic form, i.e., as designed, with little or no knowledge of the underlying data structure of the unimputed file or the subject matter content. Both were able to create an imputed file within hours of receiving the unimputed file and the set of edits. In practice, the specific file of edits and the set up of the systems can be done before the data to be edited are available.

8. SUMMARY

GEIS is to be preferred due to the generality of its algorithms, the quality of its documentation, the number of imputation options, and the fact that it is specifically designed for third party users. Because of the close performance of both systems, the practical choice between the two will likely be made based on more pragmatic reasons such as the computing environment that is used and the availability of programmers for auxiliary tasks.

ACKNOWLEDGEMENTS AND DISCLAIMER

The authors are grateful for the help provided by Pierre Caron of Statistics Canada in generating the population and running the GEIS programs. The opinions of the second author are his own and do not necessarily reflect those of the U.S. Bureau of the Census.

REFERENCES

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

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

Cotton, C. (1991), "Generalized Edit and Imputation System Functional Description," Statistics Canada Technical Report.

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

GEIS Development Team (1990), "Generalized Edit and Imputation System Specifications," Statistics Canada Technical Report.

Greenberg, B. G., Draper, L., and Petkunas, T., "On-Line Capabilities of SPEER," in *Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, 235-244.

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

Greenberg, B. G., and Petkunas, T. (1990), "Overview of the SPEER System," SRD report RR-90/15, U.S. Bureau of the Census, Washington, D.C., USA.

Kovar, J.G., MacMillan, J.H. and Whitridge, P. (1991), "Overview and Strategy for the Generalized Edit and Imputation System," Statistics Canada, Methodology Branch Working Paper BSMD 88-007E (updated in 1991).

Kovar, J.G. and Whitridge, P. (1990), "Generalized Edit and Imputation System: Overview and Applications," *Revista Brasileira de Estadistica*, **51**, 85-100, Rio de Janeiro.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, J. Wiley: New York.

Pritzker, L., Ogus, J., and Hansen, M. H. (1965), "Computer Editing Methods--Some Applications and Results," *Bulletin of the International Statistical Institute, Proceedings of the 35th Session*, Belgrade, 395-417.

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

Schiopu-Kratina, I. And Kovar, J.G. (1989), "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Statistics Canada, Methodology Branch Working Paper BSMD 89-001E.

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

Winkler, W. E. (1996), "SPEER Edit System," computer system and unpublished documentation, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA.

Table 1: Means, mean differences and MAD's (away from the clean file) for the response variables, expressed both in absolute terms ($) and, in parentheses, relative increase with respect to the means of the clean file (%).

Variable	Clean data	Unimputed data		SPEER imputed data			GEIS imputed data		
		Clean records only	All nonmiss. records	Mean	Difference	MAD	Mean	Difference	MAD
income	70194	79384 (13.1%)	335738 (378%)	70101	-93 (-0.13%)	1327 (1.89%)	69848	-346 (-0.49%)	1885 (2.69%)
inc_p1	34237	38722 (13.1%)		33855	-382 (-1.12%)	1697 (4.96%)	33393	-844 (-2.47%)	2121 (6.20%)
inc_p2	27957	31248 (11.8%)		27850	-107 (-0.38%)	1449 (5.18%)	28144	187 (0.67%)	1861 (6.66%)
inc_p3	8000	9414 (17.8%)		8395	395 (4.94%)	978 (12.2%)	8311	311 (3.89%)	997 (12.5%)
expenses	51298	57786 (12.6%)	139407 (172%)	51001	-297 (-0.58%)	1021 (1.99%)	51457	159 (0.31%)	1348 (2.63%)
exp_p1	13025	14922 (14.6%)		13120	95 (0.72%)	362 (2.78%)	13157	132 (1.01%)	495 (3.80%)
exp_p2	12482	13907 (11.4%)		12407	-75 (-0.60%)	581 (4.66%)	12486	4 (0.03%)	771 (6.18%)
assets	347950	380535 (9.4%)	348016 (0.02%)	345343	-2607 (-0.75%)	6767 (1.94%)	345928	-2022 (-0.58%)	9501 (2.73%)

Table 2: Correlations between key variables, before and after imputation, and, in parentheses, relative difference between the pre- and post-imputed values (%)

Variable pair	Clean data	Unimputed data		SPEER imputed data	GEIS imputed data
		Clean records only	All nonmissing records		
land - income	0.8354	0.8342	-0.0081	0.8354 (+0.00%)	0.8340 (-0.17%)
land - expenses	0.7965	0.7981	-0.0227	0.7977 (+0.15%)	0.7905 (-0.75%)
land - assets	0.6068	0.6110	0.6117	0.6171 (+1.70%)	0.6089 (+0.35%)
income - expenses	0.9594	0.9598	-0.0047	0.9556 (-0.40%)	0.9545 (-0.51%)
income - assets	0.7059	0.7252	0.0114	0.7119 (+0.84%)	0.7062 (+0.04%)
expenses - assets	0.6824	0.6989	-0.0012	0.6836 (+0.17%)	0.6750 (-1.08%)