ADAPTING THE SPEER EDIT SYSTEM TO EDIT HOG DATA IN THE NATIONAL AGRICULTURAL STATISTICS SERVICE'S QUARTERLY AGRICULTURAL SURVEYS

Todd A. Todaro, National Agricultural Statistics Service
USDA/NASS, Research Division, 3251 Old Lee Hwy., Room 305, Fairfax VA 22030-1504

Key Words: Automatic Edit and Imputation System, Error Localization, Fellegi-Holt, SPEER

1. INTRODUCTION

Data editing costs can significantly contribute to total survey costs, both in terms of staff time and dollars. Based on a voluntary reporting in 1990 of Federal Statistical Agencies, the Subcommittee on Data Editing in Federal Statistical Agencies (Hanuschak et al., 1990) reported that the modal cost of editing in all Federal surveys was 10 percent of the total survey costs, and the median was 35 percent of the total survey costs. These reported costs of editing warrant consideration of more efficient ways of editing the data.

The National Agricultural Statistics Service (NASS) conducts a wide variety of agricultural surveys. Among these are the Quarterly Agricultural Surveys which are used to collect current agricultural production data. Data collection is initiated at the beginning of each quarter; editing of the data must be near completion in the following two weeks; and the results are published within a month of the start of data collection. Thus, timeliness is a key issue. Since NASS must conform to a rigid schedule of collecting, editing, and publishing survey data, new and innovative procedures are sought to improve the efficiency while maintaining the timeliness of the editing process.

In August 1996, the Research Division of the National Agricultural Statistics Service began the review of an automatic edit and imputation system developed at the Bureau of the Census called "Structured Programs for Economic Editing and Referrals" (SPEER, Greenberg and Surdi (1984), Greenberg and Petkunas (1990), Winkler (1996)). The SPEER edit system is designed to edit continuous data with the edits specified either as ratio or balance edits. A ratio edit is of the form $L_{ij} \le V_i/V_j \le U_{ij}$, where V_i is the variable in the numerator of the ratio, V_j is the variable in the denominator of the ratio, L_{ij} is the lower edit bound, and U_{ij} is the upper edit bound. A balance edit is of the form $\sum_i V_i = V_s$ $(i \neq s)$, where the values of the variables V_i are required to sum to the value of the variable V_s. These edits are said to fail if the above conditions are unsatisfied when substituting variable values. If any edits fail, the SPEER edit system identifies a subset of variable values to delete and impute so that all edits are satisfied. Thus, human intervention in the edit process is minimized, once the edits are specified.

The SPEER edit system follows the Fellegi-Holt philosophy of editing. In their landmark paper "A Systematic Approach to Automatic Edit and Imputation," Fellegi and Holt (1976) discuss an automatic edit and imputation system with the following three criteria:

1. The data in each record should be made to satisfy all edits by changing the fewest possible items of data.

2. As far as possible, the frequency structure of the data file should be maintained.

3. Imputation rules should be derived from the corresponding edit rules, without explicit specification.

In addition to possessing the features of a Fellegi-Holt system, the SPEER edit system is attractive for the following two reasons. First, the editing process is repeatable (reproducible) in time and space. That is, the results of data records run through the SPEER edit system will be the same regardless of when and where they are edited. On the other hand, manual editing performed by the same or different people will not always be repeatable. Second, the SPEER edit system provides an audit trail, which is a tracking of changes made to the data records and the reasons why the changes were made. It allows for the assessment of the impact of editing and imputation on data records and their expansions. It also provides feedback that may be useful in improving future surveys.

The intent of this research effort is to determine whether the Bureau of the Census's SPEER edit system would be useful in NASS's Agricultural Survey processing, and if so, what balance of the SPEER edit system and the current edit system would be optimal. The current edit system is comprised of an interactive edit system for micro-level (record level) edits and an interactive edit system for macro-level (aggregate level) edits. Based on some preliminary analyses, the Research Division decided to adapt the SPEER edit system to edit data from its September 1996 Iowa quarterly Hog Survey and compare results of data expansions to those that are obtained using the current edit system for paper collected data.

A detailed description of the SPEER edit system is provided in Section two. Section three discusses some results of this study. Conclusions are provided in Section four.

2. THE SPEER EDIT SYSTEM

The SPEER edit system is comprised of three computer programs written using the FORTRAN programming language. The system is completely portable from one operating system to another. The output from two of the programs is used as input into the main edit and imputation program (spr3d.for). The first of these two programs (cmpbeta3.for) computes ratio regression coefficients to be used in the imputation process. The second program (gb3.for) logically derives implied ratio edits, termed implicit ratio edits, from the explicitly user-specified ratio edits, termed explicit ratio edits. The data set to be edited is read by the program spr3d.for. This data set (as well as all others) should be in ASCII format. Modifications to the SPEER edit system to edit data from different surveys are made by modifying FORTRAN format and parameter statements (which are used to specify the maximum number of variables and/or records) and input files read by the three computer programs. These input files contain input and output filenames, formats of data sets, and the number of variables included in the data sets. The SPEER edit system edits continuous data and presumes that a considerable amount of pre-editing has been done.

The SPEER edit system creates three output files. The output file "spr.out" contains data record summaries. These summaries show which edits failed, the variables with deleted values, the imputation ranges, the method of imputation, the reported variable values, and the imputed variable values. The output file "sum.out" is a summary file containing information on the number of records run through the SPEER edit system, the frequency of failed ratio edits and the frequency of deleted variable values. The output file "edit.out" is an ASCII file containing the edited data file which can be used for summarization purposes.

The SPEER edit system essentially performs three functions:

1. Editing
2. Error Localization
3. Imputation

2.1 EDITING

Prior to subjecting the data to the edits, the edits must be specified and analyzed. The specification of the ratio edits can be further sub-divided into two steps. The first step is the determination of pairs of variables that are logically related and highly correlated to form ratio edits. The second step is the specification of the lower and upper edit bounds for the ratio edits. These bounds could be determined via subject matter specialists or statistically. In this paper the current edit system uses edit bounds developed solely by subject matter specialists, while the SPEER edit system uses edit bounds developed statistically using a method called "Resistant Fences," described by Thompson and Sigman (1996). In addition to the specification of ratio edits, simple balance edits may be specified in a file used as input into the main SPEER edit and imputation program. The adjective "simple" is used to mean that any variable can be included in at most one balance edit. General balance edits are not supported by the existing algorithms in the SPEER edit system.

Once all the edits have been explicitly specified, implied edits are logically derived. The SPEER edit system derives implicit ratio edits, and induced edits from ratio and balance edits. These implied edits are vital for edit analysis. They provide feedback on the consequences of the explicitly specified edits on the data. Analyzing implied edits are useful in determining redundant edits and edits that are too

restrictive (or not restrictive enough). Implied edits are also useful in solving the error localization problem discussed in the next section. Fellegi and Holt (1976) showed that the error localization problem could be solved by generating a complete set of edits (i.e., deriving all implied edits).

2.2 ERROR LOCALIZATION

Error localization refers to identifying and deleting a (weighted) minimal set of variable values that are to be subsequently imputed so that all edits are satisfied. Note that the use of the adjective "minimal" is in conformance with Fellegi and Holt's first criterion in Section 1. The SPEER edit system allows the option to specify variable weights, W_i. A higher weight for a particular variable signifies the placement of higher confidence in the reported value of the variable, and the less likely the value of the variable will be deleted. The default weight is 1.0.

A modification was made to the SPEER edit system to avoid deleting and imputing positive values for variables that had zero reported values. Without these modifications, many variable values could have positive values imputed when the value is most likely zero.

The SPEER edit system algorithms do not generate a complete set of edits. All implicit ratio edits are generated, but not all induced edits are generated. Thus, the error localization problem may not always be correctly solved. The ramification is that the minimal set of variable values identified is not enough; more variable values need to be deleted so that all edits can be satisfied.

Draper and Winkler (1997) list as one of their five goals of the SPEER edit system that all edits be satisfied after one pass through the data. However, with the existing SPEER edit system algorithms, one pass may not always correct all data records. They suggest making the process iterative by performing multiple passes. This has also been done by NASS, using up to six iterations. Although more data records may be corrected, the only way to have all of the data records corrected is to generate a complete set of edits or to take a different approach to the solution of the error localization problem (Schiopu-Kratina and Kovar, 1989).

2.3. IMPUTATION

The SPEER edit system imputes variable values that have been deleted for a data record based on other variable values in the same data record that have not been deleted. The deleted variable values are imputed in ascending variable number order. The variable numbers are assigned in ascending order beginning with one to the number of variables on the data set to be edited as the SPEER edit system (the program spr3d.for) reads the data.

The SPEER edit system begins by attempting to impute variable values that must have a unique value given the remaining non-deleted variable values to satisfy edits. This is done by examining each balance edit to see if one has a single variable value deleted. If there is such a balance edit, then there exists a single (deterministic) value for the deleted variable such that the balance edit is satisfied. The next step is to calculate an imputation range for each remaining variable with a deleted value. An imputation range for a variable is the set of values such that any value imputed within this set in conjunction with the undeleted variable values will satisfy the edits involving these variables.

The SPEER edit system initially calculates an imputation range using only the complete set of ratio edits (all explicit and implicit ratio edits). It then attempts to further restrict this imputation range by using the failed induced edits (Kovar and Winkler, 1996). The SPEER edit system will not impute a value for a variable if an imputation range cannot be calculated.

Once an imputation range has been established, the SPEER edit system employs a hierarchical imputation scheme. The first imputation method attempted is ratio regression imputation of the form $V_i=\beta_{ij}V_j$, where V_i and V_j are the variables involved in a ratio edit in the complete set of ratio edits. If the imputed value for V_i lies outside the imputation range, then a default imputation scheme based on the imputation range is used. After each iteration, most of the variables were rounded since they are discrete in nature. If for any data record the rounding resulted in any failed edits, the data record was subjected to another iteration, up to a maximum of six. If after six

iterations, a data record was not corrected, it was then written to an output file where it could be manually inspected.

The code for imputation in the SPEER edit system can be easily modified to accommodate other imputation schemes since it is contained in a separate module of the program (i.e., a FORTRAN function). The ratio regression imputation scheme is provided because of its accuracy in economic surveys for which the SPEER edit system is primarily used at the Bureau of the Census. Giles and Patrick (1986) describe several alternative imputation estimators.

3. RESULTS FOR HOGS

Aggregate statistics from the SPEER edit system are compared to those from the current edit system for paper collected questionnaires from the September 1996 Iowa Hog Survey. Four balance edits were excluded in this study. Including these balance edits would violate the simple balance edit restriction imposed by the SPEER edit system. This resulted in thirteen records that required manual editing out of approximately 1100 records. In addition, one record was not corrected after six iterations, and thus required manual editing. These fourteen records, which SPEER could not handle, were excluded from the following results.

Of the 23 variables in this study, 13 had absolute expanded differences between the two edit systems of less than one percent. However, 6 variables (feeder pigs purchased, feeder pig price, feeder pig weight, total hogs & pigs, market hogs & pigs over 180 pounds, and boars & young males for breeding) had absolute expanded differences exceeding 10 percent as shown in Table 1.

Table 1. Comparison of Expanded Totals

Variable	SPEER Edits	Current Procedures	Percentage Difference
Feeder Pigs Purchased	224,372	180,547	24.27
Feeder Pig Price	19,616	17,144	14.42
Feeder Pig Weight	21,538	19,043	13.10
Total Hogs & Pigs	8,109,731	7,107,931	14.09
Market Hogs & Pigs Over 180 Pounds	2,343,304	1,320,620	77.44
Boars & Young Males for Breeding	32,935	27,907	18.02

The explanation for the large discrepancies for the total hogs & pigs and market hogs & pigs over 180 pounds was the result of a single record with a key entry error where the leading number was duplicated so that about one million too many market hogs & pigs over 180 pounds were entered. The SPEER edit system changed the value of the total hogs & pigs variable rather than changing the value of the market hogs & pigs over 180 pounds variable. A balance edit was specified where the sum of the market and breeding hog variable values equals the total hogs & pigs variable value. Whenever the sum of the market and breeding hog variable values exceeded the total hogs & pigs variable value, as was the case for the above mentioned record, the SPEER edit system was programmed to replace the value of the total hogs & pigs variable value with the sum of the market and breeding hog variable values. Without this record, the absolute percentage difference for total hogs & pigs would be about 0.10 and for market hogs & pigs over 180 pounds about 1.05.

It is interesting to point out the large differences between the two systems for the feeder pig variables -- feeder pigs purchased, feeder pig price, and feeder pig weight. Whenever the value of the average feeder

pig weight per head was in the range of 10 to 15 pounds, the values for all three feeder pig variables were edited and assigned zero values in the current system. This occurred 15 times. Whenever this occurred in the SPEER edit system, the ratio of average feeder pig price per head to average feeder pig weight per head exceeded the associated upper ratio edit bound. A higher variable weight, W_i, was assigned to the average feeder pig price per head variable so that the value of the average feeder pig weight per head variable would be deleted and imputed. By imputing the value for the average feeder pig weight per head variable, the resulting values for the average feeder pig price per head variable and average feeder pig weight per head variable were closer to the values for these two variables in other records. The SPEER edit system increased the value of the average feeder pig weight per head using the relationship between the average feeder pig weight per head and the average feeder pig price per head.

The large percentage difference for boars & young males for breeding was the result of the SPEER edit system changing the boars & young males for breeding variable value for five records. There were no changes made to the values of this variable in the current edit system. For two of the five records, the variable values were changed to satisfy the failed balance edit: the sum of the market and breeding hog variable values equals the total hogs & pigs variable value. The variable values for the other three records were changed to satisfy a failed induced edit. For these three records, one or both of the ratio edits involving the sows, gilts, and young gilts for breeding and the sows expected to farrow failed. The failed induced edit(s) was derived from the failed ratio edit(s) and the above balance edit.

Table 2 provides information on the amount of editing performed by the two systems. It shows the frequency of records that were 1) not changed in either the SPEER edit system or in the current edit system, 2) not changed in the SPEER edit system but changed in the current edit system, 3) changed in the SPEER edit system but not changed in the current edit system, and 4) changed in the SPEER edit system and changed in the current edit system.

Table 2. Comparisons of Same Record Changes

Variable	SPEER: No change Current: No change		SPEER: No change Current: Change	SPEER: Change Current: No change	SPEER: Change Current: Change
	Value=0	Value>0			
Total Hogs & Pigs	289	779	7	4	5
Market Hogs & Pigs under 60 LBS.	445	630	7	2	0
Market Hogs & Pigs 60-119 LBS.	427	651	5	0	1
Market Hogs & Pigs 120-179 LBS.	437	638	5	4	0
Market Hogs & Pigs over 180 LBS.	415	654	12	3	0
Boars & Young Males for Breeding	530	549	0	5	0
Sum of all variables (includes variables not listed above)	14964	9773	107	62	26

The entries in Table 2 reveal that the two systems usually did not make changes to the same record variable values. The current edit system made about 1.5 times as many variable value changes as the SPEER edit system. This is in concert with the first criterion of Fellegi-Holt systems listed in Section 1. For the feeder pig variables, the current edit system changed 53 values compared to the SPEER edit system changing only 19 values. However, the

expanded values for many of these variables are similar.

4. CONCLUSIONS

Using the SPEER edit system has several potential advantages for NASS surveys:

With the exception of a relatively small number of records, the SPEER edit system creates an edited data set similar to that currently produced by NASS. Only twenty-one records accounted for the large absolute differences in expansions.

The system is very fast. Running 1155 hog data records through the system (with a maximum of six iterations per record) took approximately 67 seconds on a Pentium 90 MHz computer.

However, there are also disadvantages to using the SPEER edit system for NASS surveys:

The system cannot handle all commodity records because of the restrictions on the specified edits (ratio and simple balance edits). Four balance edits had to be excluded when editing the hog data set because variables were in more than one balance edit. Pre-SPEER edits (e.g., editing categorical variables) and post-SPEER edits (corrections for the records the SPEER edit system could not handle) must be performed outside of the system.

Thus, the current SPEER edit system is only useful to NASS when all edits can be specified as ratio edits. The edits in NASS surveys are more complex and generally require many balance edits that violate the simple balance edit restriction in the current SPEER edit system.

REFERENCES

Anderson, C. et al. (1996), "Report of the Hog Editing and Analysis Team," unpublished documentation, U.S. National Agricultural Statistics Service, Washington, D.C.

Draper, L.R., and Winkler, W.E. (1997), "Balancing and Ratio Editing with the New SPEER System," Proceedings of the Survey Research Section, to appear.

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

Giles, P., and Patrick, C. (1986), "Imputation Options in a Generalized Edit and Imputation System," Survey Methodology, Vol. 12, No. 1, 49-60.

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

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

Hanuschak et al., (1990), Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology, Data Editing in Federal Statistical Agencies, Statistical Policy Working Paper 18.

Kovar, J.G., and Winkler, W.E. (1996), "Editing Economic Data," Proceedings of the Survey Research Section, 81-87.

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 No. BSMD-89-001E.

Thompson, K.J., and Sigman, R.S. (1996), "Statistical Methods for Developing Ratio Edit Tolerances for Economic Censuses," Proceedings of the Survey Research Section, 166-171.

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