United States Department of Agriculture

National Agricultural Statistics Service

Research Division

RD Research Report Number RD-97-04

November 1997

Evaluation of the SPEER Automatic Edit and Imputation System

Todd A. Todaro

EVALUATION OF THE SPEER AUTOMATIC EDIT AND IMPUTATION SYSTEM, by Todd A. Todaro, Technology Research Section, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250-2000, November 1997, NASS Research Report, Report No. RD-97-04

ABSTRACT

Data editing plays an important role in the survey process. The National Agricultural Statistics Service currently uses, in addition to some manual editing, an interactive micro-level edit system and an interactive macro-level edit system to edit reported data. Advantages of using these two edit systems are that: 1) the most complex edits can be incorporated and 2) the impact of editing at aggregate levels can be readily evaluated. There are, however, disadvantages with the use of the two edit systems: 1) a considerable amount of time and resources may be expended and 2) editing may not always be performed in a similar, consistent manner within a State Statistical Office (SSO) or between SSO's.

This paper evaluates an automatic edit and imputation system developed at the Bureau of the Census called Structured Programs for Economic Editing and Referrals (SPEER). The SPEER edit system is appealing for the following reasons: 1) the editing process is fully automated (editing and imputation is performed by the system), 2) the results of data run through the SPEER edit system are reproducible in time and space, and 3) a minimal amount of editing is performed so that all edits are satisfied. Comparisons between the SPEER edit system and the current edit procedures are made for expanded totals and the number and magnitude of variable value changes. The data used for these comparisons are obtained from the Quarterly Hog Survey and the Quarterly Agricultural Survey. Results of the comparisons reveal that a number of data records run through the SPEER edit system could not be made internally consistent due to limitations on the specification of the edits. Implementation of the current version of SPEER is not recommended. Future edit and imputation alternatives are also presented.

KEY WORDS

Automatic Edit and Imputation System; SPEER; Error Localization; Resistant Fences

ACKNOWLEDGMENTS

The author would like to thank the Iowa SSO and North Carolina SSO for providing Key-Entry III data files for use in this study; Dale Atkinson and Roberta Pense for helpful guidance throughout the project and for reviewing drafts of this report; and William E. Winkler of the Bureau of the Census for providing the SPEER computer programs. Thanks to management (Ron Bosecker and George Hanuschak) for their support of this project.

TABLE OF CONTENTS

SUMMARY

The National Agricultural Statistics Service (NASS) collects and summarizes information about the nation's agriculture through the use of a variety of surveys and the upcoming 1997 Census of Agriculture. After data collection and prior to the summarization and publication of statistics, the data are edited for completeness and consistency. Obtaining edited data that are accurate is important for making inference of the underlying population characteristics (e.g., estimating population totals and ratios). Edited data are also used as control data for designing future surveys and improving the accuracy of the estimates from them.

It is desirable for the editing process to be efficient and expeditious. NASS currently collects data via two primary modes -- Paper questionnaires and Computer Assisted Telephone Interviewing (CATI). For the paper mode, the data are edited manually and also with micro- and macro-level machine edits. For the CATI mode, the data are edited using micro- and macro-level edit systems. The current editing process can often be time consuming, requiring considerable staff hours to complete the editing tasks. Hence, the data editing costs can make up a noticeable portion of the total survey cost. Moreover, NASS surveys must often be completed under tight time constraints. For example in the Quarterly Agricultural Surveys, data collection is initiated near the beginning of the month for each quarter; editing of the data must be near completion in the following two weeks; and the survey results are published at the end of the month. Since NASS must conform to a rigid schedule of collecting, editing, and publishing survey data, new procedures are constantly sought to improve the editing process.

The Structured Programs for Economic Editing and Referrals (SPEER) edit system offers the potential to improve the efficiency of the editing process while also performing editing in a timely manner. The SPEER edit system is designed to edit continuous data with the edits specified as ratio edits or balance edits. It is based on the Fellegi-Holt model of editing (Fellegi and Holt, 1976) which has the following three criteria:

1) The data in each record should satisfy all edits by changing the fewest possible variable values.
2) As far as possible, the frequency structure of the data file should be maintained.
3) The imputation rules should be derived from the corresponding edit rules without explicit specification.

For data records failing one or more edits, the minimal set of variable values is identified to be deleted and subsequently imputed, so that all edits are satisfied. There are, however, some limitations associated with the use of the SPEER edit system. The edits must be formulated either as ratio edits 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. In addition, the balance edits are further restricted in that a variable can be involved in at most one balance edit. Another limitation is that the SPEER edit system does not always identify the minimal set of variable values to change. Fellegi and Holt (1976) showed that this minimal set could be found by logically deriving all implied edits

from the edits explicitly specified. But, the SPEER edit system does not generate all implied edits. As a result, a data record may not be corrected after one pass through the SPEER edit system.

The results of this study reveal that the SPEER edit system failed to correct approximately 20 percent of the records that it attempted to correct for the data from the Quarterly Hog Survey. Nearly all of these could not be corrected because of the restriction on the balance edits. In addition, data records for which the difference between the imputed and originally reported variable values was large, would likely be manually reviewed. Thus, a large number of data records that the SPEER edit system modified would still need to be manually reviewed. The results of adapting the SPEER edit system to edit data from the Quarterly Agricultural Survey were no more promising. Many data records in which a change was made by the SPEER edit system were not corrected, and thus would require editing by another means. Since the SPEER edit system would not eliminate any of the components of the current editing process, it is recommended that the SPEER edit system not be implemented in NASS's Agricultural Survey processing in its current form.

Several future alternatives are also presented briefly, such as profile editing using the Data Warehouse, and resistant fences to specify edit bounds for current NASS edit systems.

1. INTRODUCTION

Data editing costs can significantly contribute to the 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 cost, and the median was 35 percent of the total survey cost. This reported cost of editing warrants 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 Quarterly Agricultural Surveys which are used to collect current agricultural production data. Data collection begins around the first of each quarter; editing of the data must be near completion in the following two weeks; and the results are published at the end of the month. Thus, timeliness is an important attribute of the quality of the data. Currently, a CATI instrument is used to collect data whenever possible. However, some data are also collected via paper questionnaires. 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 (NASS/RD) 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. If the SPEER edit system is unable to impute values such that all edits are satisfied, then the data record is written to a file where it can be manually inspected. (This is a modification made by NASS.) Thus, human intervention in the edit process is minimized, once the edits are specified.

The SPEER edit system as well as other automatic edit and imputation systems, most notably, GEIS (Generalized Edit and Imputation System) developed at Statistics Canada (see for example, Kovar et al., 1991), follow 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.

The first criterion attempts to preserve as much of the originally reported data as possible, under the premise that errors in the data fields occur rarely. The aim of the second criterion is to maintain the marginal and joint frequency distributions of the clean data (i.e., those data initially satisfying all edits). Thus, when data must be manufactured via imputation, it is desirable for the frequency distributions to remain unchanged. The third criterion ensures that the resulting partially imputed data will satisfy the specified edits. This criterion is attractive in that it avoids generating imputed data that do not meet the standards set for reported data.

It is noted that Fellegi-Holt systems can handle linear edits, but not nonlinear edits. Some conditional edits can be handled by reformulating them into ratio edits. (This is illustrated in Appendix B.)

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 were 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 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. Based on some preliminary analyses, the Research Division decided to adapt the SPEER edit system to edit data from its Quarterly Hog Survey and compare results of data expansions and indications to those that are obtained using the current editing system. The primary editing tools in the current system are the Blaise interactive edit system for micro-level edits (record level) and the Interactive Data Analysis System (IDAS) for macro-level edits (aggregate level). The current NASS systems do not necessarily protect the frequency or correlation structure of the data. The primary reason for choosing data from the hog survey was that the set of edits had been recently updated in the Hog Edit and Analysis Team (HEAT) report (see Anderson et. al., 1996).

Key-Entry III files for paper collected reports from the September 1996 Iowa Hog Survey were used in this study. These data had been manually, but not machine, edited. Inaccessible and refusal nonrespondents, as verified against the final survey data file, were removed from the data set. Duplicate records in the key-entry files, with the same values for the ID, tract and subtract variables, were eliminated. This was done by comparing the duplicate records with the corresponding records in the final survey file. The result of these comparisons was to either keep one of the duplicate records or combine the duplicate records into a single record. A few of the data records indicated as refusals in the key-entry files contained positive values for survey variables. These records were also compared with the final survey file for their possible inclusion.

Section 2 of this paper discusses previous studies that have compared the SPEER edit system with other edit and imputation

systems. A detailed description of the SPEER edit system is provided in Section 3. Section 4 discusses some results of this study. Section 5 discusses some intricacies and problems of adapting the SPEER edit system to the December 1996 Agricultural Survey. Conclusions and recommendations are provided in Section 6.

2. LITERATURE REVIEW

Pierzchala (1990) provides an overview of three edit and imputation systems: Blaise, GEIS, and SPEER. Prior to discussing each system, he lists four questions to aid an organization in selecting an editing system.

They are:

1. Which kinds of data and surveys must be handled?
2. Where must it fit into the survey processing flow?
3. Which environments must it operate in?
4. Which kinds of edits can it handle?

Tables are provided by the author for the three systems as to how they address each of the above questions. The similarities/differences of the three systems can be ascribed to their designed purpose in the editing process.

Pierzchala points out a distinction among the three systems. Blaise is referred to as an integration system while GEIS and SPEER are referred to as Fellegi-Holt systems. An integration system is defined as one that "seeks to perform as many different survey functions as possible within one system." These survey functions include data collection, data entry, and editing. Fellegi-Holt systems are primarily concerned with micro-editing after data collection. These systems presume that a considerable amount of pre-editing has been done. Examples of pre-edits include checking that all response coding in the record is logically consistent. Those data records that remain unresolved or that have a minor impact on aggregate statistics can then be run through a Fellegi-Holt system to resolve any persistent inconsistencies.

Kovar and Winkler (1996) provide a comparison of SPEER and GEIS using simulated data intended to resemble a portion of data from the Canadian Agriculture Survey. This simulated data set was then perturbed to create a data file that was edited using both systems. The specified edits were nearly equivalent for the two systems, differing only because of the restriction on the formulation of edits (ratio or simple balance edits) in the SPEER edit system. The two systems were compared in terms of their ability to restore mean variable values in the perturbed data file to the corresponding means of the simulated data file and preserve the correlation structure of the variables. The systems were also compared on ease of use, computational speed, and generality.

Both systems were similar in restoring the means and preserving the correlation structure of the variables. Neither system was preferred on ease of use. The SPEER edit system processed the data file faster due to the simplicity of the algorithms used. GEIS was preferred on the basis of generality since it handles general linear edits, whereas the SPEER edit system handles ratio and balance edits which are subsets of general linear edits. Overall, GEIS was the preferred system since it was deemed more general, provided more imputation options, and was well documented. (The SPEER edit system was primarily intended for internal use in the Bureau of the Census.)

As an aside, GEIS was written in the C programming language with frequent subroutine calls to the ORACLE database. The primary reason for not considering the evaluation of GEIS is that NASS does not use ORACLE as its database platform.

Imel and Ramos (1992) of the Bureau of the Census, conducted two studies which compared an adapted SPEER edit system (Ag-SPEER) and the Ag-Complex edit and imputation system, which has historically been used to edit the Census of Agriculture. Ag-Complex contains modules that perform historical and consistency checks. The historical edit module compares data records with the same census file number from the current census to the previous census. This module identifies extreme economic size and type of farm differences. The consistency module is designed to ensure within record consistency in the data items. These studies evaluated the performance of Ag-Complex relative to Ag-SPEER, with the second study a follow-up to the first.

Both studies used unedited crop and land data for thirty to forty variables from the 1987 Agricultural Census. A "clean" data set was identified as those records which satisfied all edits from both systems. The two studies differed on how the clean data records were perturbed (methodology and number of variables on each record) and whether or not a subjective evaluation was included. In the first study, all variables on each record were modified. In the second study, the clean data set was replicated so that the number of replicates equaled the number of variables on the data set. Then, a single variable was modified on each record. The two studies were compared by using the average absolute bias of the edited values, and by using a mean square error (MSE) measure. The mean square error measure was calculated using squared deviations of the edited data about the clean mean of the variable in the first study, and about the clean data values in the second study. The second study also included a subjective evaluation component.

Based on the above two criteria measures, the first study concluded that the Ag-Complex system outperformed the Ag-SPEER system. However, since the assumption that all variables in a single record would be perturbed was deemed questionable (by those performing the study), the second study was planned. In the second study, when the replicated files were combined both systems had an equal number of variables with the lowest MSE measure; and Ag-SPEER had slightly more variables with a lower average absolute bias. Based on these results, it seems reasonable to conclude that the Ag-SPEER system is at least as good as the Ag-Complex system. However, after considering the subjective evaluation component, it was concluded that the Ag-Complex system was preferred to the Ag-SPEER system.

The subjective evaluation consisted of subjectively comparing edited output from the Ag-SPEER and Ag-Complex systems for two files of 1,000 records each, termed analysis file 1 and analysis file 2. Three analysts were asked to identify for each record which edited output was the best for both analysis files. The results are presented in Table 1 and Table 2.

Table 1. Number of Records Considered Best for Analysis File 1[1]

	Analyst1	Analyst2	Analyst3
Complex	184	102	161
SPEER	91	102	79
Tie	725	796	760

Table 2. Number of Records Considered Best for Analysis File 2[1]

	Analyst1	Analyst2	Analyst3
Complex	197	134	206
SPEER	88	141	64
Tie	715	725	730

Analysts 1 and 3 selected more records from Ag-Complex as producing the best edited record. The second analyst selected about equal numbers of records from both systems as producing the best edited record. However, for 70 to 80 percent of the records, the analysts could not distinguish the best output from the two systems. Tables 3 and 4 were created based on Tables 1 and 2, respectively. The entries in these tables were created by dividing the number of times an edit system was selected as producing the best output plus the number of ties by the total number of records.

Table 3. Proportion of Records Considered as Good as or Better for Analysis File[1]

	Analyst1	Analyst2	Analyst3
Complex	0.909	0.898	0.921
SPEER	0.816	0.898	0.839

Table 4. Proportion of Records Considered as Good as or Better for Analysis File 2[1]

	Analyst1	Analyst2	Analyst3
Complex	0.912	0.859	0.936
SPEER	0.803	0.866	0.794

Second, the conclusion of the authors connotes that the systems corrected the perturbed value (i.e., the perturbed value of the variable was replaced with the original, clean value) which is misleading. The analysts were simply presented with the perturbed variable values and the edited data values and asked which edited record appeared to be the

[1]Tables taken from Imel, J., and Ramos, M. (1992).

best. Third, since the analysts were familiar with Ag-Complex imputation, they were more likely to identify the best record as the one which was imputed in a more familiar manner (i.e., imputation using Ag-Complex). It is also interesting to point out the large differences between analysts in identifying the best edited record. This point implores the question of how the analysts were chosen and what experiences they had with Ag-Complex.

3. DETAILED DESCRIPTION OF THE SPEER EDIT SYSTEM

The SPEER edit system consists of three computer programs written using the FORTRAN programming language. The system is completely portable from one operating system to another. The outputs from two of the programs are used as inputs 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 explicit ratio edits together with all the derived implicit ratio edits form a complete set of 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 incorporated into 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 numbers 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 (e.g., editing section completion codes).

The spr3d.for program essentially performs three functions:

1. Editing
2. Error Localization
3. Imputation.

3.1 EDITING

3.1.1 EDIT SPECIFICATION

Prior to subjecting the data to the edits, the edits must be specified and analyzed. This editing step, usually iterative, is extremely important since it has a direct bearing on the final data quality. The specification of the ratio edits can be further subdivided into two steps. The first step is the determination of pairs of variables that are logically related and highly correlated to form ratio edits. This generally requires the expertise of subject matter specialists and the use of a statistical package to perform some correlation analyses of the data. 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), with some modifications. In addition to the 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. This is a major limitation of SPEER when the variables involved in the ratio edits are the sum of several variables and these variables violate the simple balance edit restriction

As a simple example illustrating this limitation, consider the following two ratio edits:

$L_{12} \leq V_1/V_2 \leq U_{12}$ where $V_1=V_5+V_6+V_7$
$L_{34} \leq V_3/V_4 \leq U_{34}$ where $V_3=V_5+V_6$

Dummy variables, V_1 and V_3 are created to formulate the above two ratio edits. Since the variable V_5 is involved in both of the above balance edits, only one can be specified in the SPEER edit system. If the first balance edit, $V_1=V_5+V_6+V_7$, is specified and the value for variable V_5 is identified to be deleted and subsequently imputed, the value for variable V_3 will not be updated, since the balance edit $V_3=V_5+V_6$ could not be specified. If the ratio edit $L_{34} \leq V_3/V_4 \leq U_{34}$ was initially satisfied, it may no longer be satisfied after imputing for V_5. But the SPEER edit system would treat the ratio edit as being satisfied.

Winkler (1996) has noted that well over 99% of the economic surveys conducted at the Bureau of the Census require variables to be involved in at most one balance edit. Thus, this limitation does not appear to be a practical problem for the Bureau of the Census. However, for the NASS Quarterly Hog Survey data set, there were: thirty variables required to formulate the edits; thirteen explicit ratio edits; and five balance edits. Four balance edits were not specified because doing so would violate the simple balance edit restriction (Refer to Appendix B). Although the December Agricultural Survey (crops and stocks) does not require as many balance edits as livestock surveys, one balance edit could not be specified due to the simple balance edit restriction.

3.1.2 RATIO EDIT BOUND DEVELOPMENT

The purpose of calculating the ratio edit bounds is to ascertain which of the ratio values are reasonable and which are not. If ratio values do not lie within the bounds, reported data will be modified, thus affecting the final, expanded estimates. Therefore, the calculation of ratio edit bounds for the SPEER edit system is very important and should be done carefully. Data from previous final survey data files are used in calculating the ratio edit bounds.

It is desirable to choose a method so that the bounds are not overly influenced when making small changes to the data (small changes to a large portion of the data, or large changes to a small portion of the data). Such a method is called a resistant method. It would certainly be undesirable for an edit bound to be unduly influenced by a few values caused, for example, by key entry errors. A resistant method is said to lose its resistance when its breakdown point is exceeded. Hoaglin et al., (1983) define the breakdown point of an estimator as "the largest possible fraction of the observations for which there is a bound on the change in the estimate when that fraction of the sample is altered without restriction."

A method called resistant fences (Hoaglin et al., 1986) is used to calculate the ratio edit bounds in this study. A SAS program written by Thompson and Sigman (1996) implements the resistant fences method, which is an exploratory data analysis outlier detection method for calculating upper and lower edit bounds for ratio edits. This method calculates the edit bounds using sample quartiles. Given an ordered set of ratios, let

q25 = the first quartile
q75 = the third quartile
H = q75 - q25, the interquartile range.

This method consists of a set of rules depending on the value of a specified constant, k. The resistant fences rules define outliers as those ratio values greater than the upper edit bound, q75+k*H, or less than the

lower edit bound, q25-k*H. For k=1.5, the rule is called the inner fence rule; for k=2.0, the rule is called the middle fence rule; for k=3.0, the rule is called the outer fence rule. The breakdown point of the resistant fences method is approximately 25% since it is based on the fourths of the sample.

Choices of various options are available to the analyst prior to running the resistant fences program. The primary options available for the creation of the ratio edit bounds are: 1) the value for k that determines which of the above resistant fences rules is used; 2) whether or not an attempt should be made to symmetrize the distribution of ratios prior to calculating the edit bounds; 3) the method of calculating the variable average (This option is useful for calculating a different value of β_{ij} for use in ratio regression imputation); and 4) the substitute value for a negative lower edit bound. Several output options include the creation of ASCII and/or SAS/GRAPH plots (SAS 1990a) and the creation of permanent SAS data sets. For the hog data, the outer fences rule was used; the data were not symmetrized prior to calculating the ratio edit bounds; the least squares method was selected for calculating β_{ij}; and zero was used as the substitute for a negative lower edit bound.

The calculation of the lower and upper edit bounds implicitly assumes a symmetric distribution of ratios since the same quantity, k*H, is subtracted from the first quartile and added to the third quartile. For data that are highly skewed, transformations can often be applied to symmetrize the data. Many data values that appeared as outliers in the untransformed data set become consistent with the rest of the distribution in the transformed data set. Thus, the edit limits can often be improved by using the symmetrizing option.

The symmetrizing option works as follows. The resistant fences program calculates edit bounds for the unsymmetrized distribution of n ratio values using the outer fences rule (i.e., k=3). To minimize the effects of multiple outliers, observations lying outside the edit limits are temporarily removed (n-n′ of them). The resulting n′ data values are now used to estimate the parameter p of the following transformation:

$$T_p(R) = \begin{cases} R^p & (p>0) \\ \log_e(R) & (p=0) \\ -R^p & (p<0) \end{cases}$$

where R represents the ratio data values.

The procedure to calculate p is a modification of the procedure described by Hoaglin, et al., (1983). Letting M be the median of the n′ ratio values and, X_L and X_U the respective lower and upper values of a set of k-percent approximate quantiles (e.g., 25th percentile and 75th percentile), order the values (X_{vi}/X_{hi}) where:

$$X_{vi}=[(X_U+X_L)/2]-M,$$
$$X_{hi}=[(X_U-M)^2+(M-X_L)^2]/[4M],$$

and set the transformation parameter p=1-median(X_{vi}/X_{hi}). The percentiles are calculated using the PCTLPTS=4 option in PROC UNIVARIATE (SAS 1992).

The number of sets of k-percent approximate quantiles used to estimate p is now addressed. A compromise must be made between the number of sets of quantiles used to estimate p and the value of the breakdown point; the larger the number of sets of quantiles used, the lower the breakdown point. To see this, consider the following example provided in Thompson and Sigman (1996). Three sets of quantiles were used: the 25th and 75th percentiles (fourths); the 12.5th and 87.5th

percentiles (eighths); and the 6.25th and 93.75th percentiles (sixteenths) with n'=130. The breakdown point is (approximately) 1/16. That is, almost 1/16 of the observations (≈8 observations) in the sample can be replaced by arbitrary values without causing the estimate of p to become unbounded. If, however, the additional set of quantiles is used, the 3.125th and 96.875th percentiles (thirty-seconds), then the breakdown point is reduced to (approximately) 1/32 (≈4 observations). Thus, with the additional set of quantiles, the breakdown point is cut in half.

Thompson and Sigman (1996) compared two methods that differed in the expected number of observations in the quantile with the smallest tail probability before the breakdown point was exceeded (E[outliers]). The selected method (E[outliers]=4) which is implemented in their program is given in the following table. Note that to avoid exceeding the breakdown point (at least in expectation), the expected number of observations in the quantile with smallest tail probability always exceeds 4. This method was selected since it resulted in the largest number of skewed data sets successfully symmetrized.

Table 5. Sets of Quantiles for EDA Transformation Plot for Symmetry

E[outliers]=4		
Range of n'	Breakdown Point	# of obs. used to estimate p
33-64	1/8	2
65-128	1/16	3
129-256	1/32	4
257-512	1/64	5
513 -∞	1/128	6

The third column, the number of observations used to estimate p, always uses as the first observation, the fourths (25th percentile and 75th percentile). The remaining observations used for any row is the set of quantiles corresponding to the breakdown point value in the row plus all sets of quantiles corresponding to breakdown point values in preceding rows.

Once the value of p is estimated, the resistance fences program calculates the skewness coefficient S (see PROC UNIVARIATE, SAS 1992) using all n ratio values for the unsymmetrized distribution of the ratios, the transformed distribution of ratios, and the natural logarithm transformed distribution of ratios. Ratio edit bounds are calculated using the transformation resulting in the smallest absolute value of S and the user-specified value of k. The final ratio edit bounds are obtained by applying the inverse of the transformation to the transformed edit bounds.

3.1.3 EDIT ANALYSIS

Once all the edits have been explicitly specified as described in Section 3.1.1, implied edits are derived. 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 for determining redundant

edits and edits that are too restrictive (or not restrictive enough). Redundant edits are edits that if removed do not further restrict the (acceptance) region defined by the edit set. Edits that are too restrictive have the value of their upper edit bound too low and/or lower edit bound too high, resulting in more failed edits than desired (analogous to the type I error).

One type of implied edit is the implicit ratio edit. As an example, consider the following two explict ratio edits:

$$1 \le V_1/V_2 \le 4 \text{ and } 4 \le V_1/V_3 \le 8.$$

An implied ratio edit, then, is: $1 \le V_2/V_3 \le 8$. The computer code for generating all implied ratio edits is relatively straightforward.

In addition to the implicit ratio edits, implied edits, termed induced edits, are also generated in the main edit and imputation program. An induced edit is logically derived from a ratio edit and a balance edit. Draper and Winkler (1997) make the distinction between simple and nonsimple induced edits. They define a simple induced edit as an "induced edit obtained by replacing only one term in a balance equation (edit) with the appropriate terms from a ratio edit." All other induced edits are nonsimple induced edits. They also identify redundant and nonredundant simple induced edits in their paper. This is helpful as only a subset of simple induced edits need to be considered. The SPEER edit system does not generate nonsimple induced edits.

As an example of the derivation of induced edits, consider the following two ratio edits and balance edit:

$$L_{14} \le V_1/V_4 \le U_{14}, \quad L_{24} \le V_2/V_4 \le U_{24}, \quad \text{and} \quad V_1 = V_2 + V_3 + V_4.$$

The first ratio edit can be rewritten as $L_{14}V_4 \le V_1 \le U_{14}V_4$. Substituting $V_2+V_3+V_4$ for V_1 yields the following simple induced edit:

$$L_{14}V_4 - V_2 - V_3 \le V_4 \le U_{14}V_4 - V_2 - V_3.$$

A nonsimple induced edit is easily obtained by substituting in the simple induced edit for V_2 using $L_{24}V_4 \le V_2 \le U_{24}V_4$. This yields:

$$L_{14}V_4 - U_{24}V_4 - V_3 \le V_4 \le U_{14}V_4 - L_{24}V_4 - V_3.$$

Note that the SPEER edit system analyzes only the complete set of ratio edits for logical consistency. It does not analyze balance edits or induced edits. Thus, careful attention should be paid when specifying balance edits. Redundant edits would not cause problems but edit bounds that are too restrictive can cause logical inconsistencies. William E. Winkler, via personal discussion, has demonstrated how linear programming techniques may be used to check the logical consistency of the combination of ratio and balance edits simultaneously using SAS (see PROC LP, SAS 1990b). Essentially, linear programming is used to determine whether or not a feasible region defined by the edit set exists. This is not a feature of the SPEER edit system, however.

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., all implied edits).

3.2 ERROR LOCALIZATION

Error localization refers to identifying and deleting a minimal (weighted) set of variable values that are to be subsequently imputed. Note that the adjective "minimal" was used to be in conformance with Fellegi and Holt's first criterion in Section 1. The manner in

which the SPEER edit system solves the error localization problem is now discussed. A counting variable bdeg(i) is created for each variable V_i, i=1,...,n in the data set. The counting variables bdeg(i) are incremented each time variable V_i is involved in a failed edit, whether it be ratio, balance, or induced, with the following modifications made by NASS. For ratio edits, bdeg(i) are incremented only if both variables involved in the ratio edit had positive reported values. For induced edits, bdeg(i) are incremented only for those variables that had positive reported values. These modifications were made to avoid 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. It is possible that an induced edit may contain a variable multiple times (say, r), in which case the corresponding counting variable bdeg(i) would be incremented by r. After accounting for all failed edits, the values of the bdeg(i) variables are tallied. These values are divided by W_i, the specified reliability-weight assigned to each variable. A higher weight for a particular variable signifies the placement of higher confidence in the reported value of the variable. The default weight is 1.0.

Two examples illustrating the usefulness of these variable weights are: 1) when a balance edit is specified, the value of the summed variable may be more reliable than the values of the component variables. Thus, the summed variable would be assigned a higher weight signifying more confidence placed on its reported value; 2) when historical variables are used in edits, their values are generally considered more reliable than the values of the survey variables they are involved with in the edits. Thus, to avoid deleting the values of historical variables, they are assigned higher weights.

The variable V_i which has the largest value of $\{bdeg(i)/W_i\}$ is the first variable to have its value deleted. The denotation of a deleted variable value in the SPEER edit system is the assignment of -1 to the variable value. In the case of ties, the SPEER edit system selects the lowest numbered variable. (The SPEER edit system numbers variables in the order that it reads the variables) Since this may lead to a biased selection of variable values selected for deletion, the code in the SPEER edit system was modified by NASS to randomly select the variable value to delete in the case of ties. Once the value of variable V_i has been deleted, the values of the bdeg(i) variables are decremented. This is accomplished by examining each failed edit involving variable V_i. For each of these failed edits, for V_j, j=1,...,n, the values of bdeg(j) are decremented by the number of times variable V_j is involved in the failed edit. After decrementing the values of bdeg(j) for all failed edits involving variable V_i, if all values of the bdeg(j) (j=1,...,n) variables are zero then the minimal set has been identified. If at least one value of bdeg(j) (j=1,...,n) is nonzero, the above process is repeated to identify the next variable value to delete. Once bdeg(j)=0 for j=1,...,n, the minimal set has been identified and the process is terminated.

The above procedure to identify the weighted minimal set of variables to delete can be formulated as a set covering problem in Operations Research for edits in the form of ratios (Greenberg, 1986).

This idea of identifying a weighted minimal set of variable values to delete originated from Fellegi and Holt (1976). They emphasized the need to generate a complete set of edits. Greenberg (1982) provides very insightful examples of why a complete set of ratio edits, and not merely the explicit ratio edits, is needed to solve the error localization problem, when all edits are in the form of ratios.

The algorithms in the SPEER edit system do not generate a complete set of edits. All implied ratio edits and all nonredundant simple induced edits are generated, but nonsimple induced edits are not 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. A particular example where the error localization is improperly solved is given in Appendix A.

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 algorithms in the SPEER edit system, 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 six iterations. Although more data records may be corrected, the only way to have all of the data records corrected is to consider a complete set of edits or to take a different approach to the solution of the error localization problem.

GEIS, mentioned earlier, approaches the solution of the error localization problem differently. Rather than using implied edits for error localization, a cardinality constrained linear program is solved using a modification of Chernikova's algorithm. (Schiopu-Kratina and Kovar, 1989). This approach, although much more computationally intensive, will correctly identify the minimal set of variable values to delete. They note that the use of implied edits for error localization is inefficient, especially when a large number of variables and edits are involved.

3.3 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 (SPEER.DAT) as SPEER (the program spr3d.for) reads the data. The SPEER edit system will not impute a value for a variable in a connected set if all other variables in the connected set have deleted or zero variable values. A set of variables is connected if, for each pair of variables, there is an explicit ratio edit involving the variables or an implied ratio edit can be derived involving the variables.

A simple example to help elucidate the definition of a connected set is now given. Consider the following explicit ratio edits:

$$L_{12} \le V_1/V_2 \le U_{12}$$
$$L_{14} \le V_1/V_4 \le U_{14}$$
$$L_{27} \le V_2/V_7 \le U_{27}$$
$$L_{35} \le V_3/V_5 \le U_{35}$$

The explicit ratio edits with the following three implicit ratio edits form a complete set of ratio edits.

$$L_{14}/U_{12} \le V_2/V_4 \le U_{14}/L_{12}$$
$$L_{12}L_{27} \le V_1/V_7 \le U_{12}U_{27}$$
$$L_{12}L_{27}/U_{14} \le V_4/V_7 \le U_{12}U_{27}/L_{14}$$

Two connected sets may be formed. The first connected set S_1 consists of the variables $\{V_1, V_2, V_4, V_7\}$ since there is a ratio edit involving each pair of variables in the complete set of edits. The second connected set S_2 consists of the variables $\{V_3, V_5\}$. The set S_2 contains only those variables involved in the fourth explicit ratio edit since no implicit ratio edit can be derived from this explicit ratio edit and another explicit ratio edit.

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. After imputing a value for this variable, however, the SPEER edit system does not always check if the resulting imputed value satisfies all edits. To illustrate this, suppose that there are only two edits, both involving V_i; one in a balance edit and one in a ratio edit, and:

$$5 \le V_i/V_j \le 10 \text{ and } V_i = V_k + V_l + V_m.$$

Let V_i=125, V_j=15, V_k=55, V_l=50, and V_m=60. The ratio edit is satisfied whereas the balance edit is not. Suppose the only variable value to be deleted and imputed is V_i. The value imputed for V_i is: $V_k + V_l + V_m = 55 + 50 + 60=165$. But the ratio edit is now no longer satisfied. If all edits for the data record are not satisfied after six iterations, the data record is written to file (check.out) where it can be manually inspected. This use of imputation can also result in inconsistent imputation ranges (discussed in the next paragraph) for variables that are remaining to be imputed. In addition, the SPEER edit system may impute a negative value for a variable in order to satisfy the balance edit. If this occurs, the code has been modified by NASS to not accept this value, but rather to impute the reported value.

The next step is to calculate imputation ranges for variables with deleted values. An imputation range for a variable is the set of values such that any imputed value within this range in conjunction with the undeleted variable values will satisfy the edits involving these variables.

For each deleted variable value, the SPEER edit system initially calculates an imputation range using only the complete set of ratio edits. To illustrate the calculation of an imputation range for a deleted variable V_i using ratio edits, consider the ratio edit $L_{ij} \le V_i/V_j \le U_{ij}$ with the value of variable V_i deleted. L_{ij} is the lower edit bound and U_{ij} is the upper edit bound for the ratio edit involving variables V_i and V_j. This ratio edit can be used only if the value for V_j has not been deleted. Several imputation ranges (say, t) for V_i could be calculated as: $V_j*L_{ij} \le V_i \le V_j*U_{ij}$ (j=1,...,t). After calculating all of these ranges, a single imputation range needs to be determined from the following t ranges for V_i:

$$V_1*L_{i1} \le V_i \le V_1*U_{i1}$$
$$\vdots$$
$$V_t*L_{it} \le V_i \le V_t*U_{it}$$

Since it is desired to find a value for V_i such that this value in conjunction with the un-deleted variable values will satisfy the edits involving these variables, the imputation range (using ratio edits) is calculated as:

$$\max_j (V_j*L_{ij}) \le V_i \le \min_j (V_j*U_{ij})$$

Now that there exists an imputation range for V_i, the SPEER edit system attempts to further restrict this imputation range by using the failed simple induced edits (Kovar and Winkler, 1996; Draper and Winkler, 1997). The SPEER edit system searches for an induced edit involving V_i where V_i is the only variable with a deleted value. If such an induced edit exists, the imputation range is calculated for variable V_i. If the lower bound of this imputation range exceeds the lower bound of the imputation range calculated from the ratio edits, then the lower bound of the imputation range for V_i is replaced by this value. Similarly, if the upper bound of the

imputation range is less than the upper bound of the imputation range calculated from the ratio edits, then the upper bound of the imputation range for V_i is replaced by this value. This process is repeated for all failed simple induced edits involving V_i, where V_i is the only variable whose value is deleted.

If after six iterations, an imputation range is unacceptable, which may occur as a result of improperly solving the error localization problem (see Appendix A), a message to this effect is printed on the individual data record and the data record is output to a file (check.out) where it can be manually inspected.

Once the 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. The resistant fences program used to calculate statistical ratio edit bounds from historical data is also used to calculate ratio regression coefficients while performing outlier analysis. (There are two advantages of using the resistant fences program over the cmpbeta3.for program. The first is that the resistant fences program will eliminate outliers in the calculation of the coefficient. The second is that the resistant fences program provides more options for calculating the coefficient). This program allows for the calculation of several different types coefficients:

(1) Ratio regression coefficients only for pairs of variables whose ratio is within the tolerances, $\sum x_i y_i / \sum x_i^2$, where $[a \le (y_i/x_i) \le b]$

(2) The median of the distribution of ratios within tolerances, or median(y_i/x_i), where $[a \le (y_i/x_i) \le b]$

(3) The mean of the distribution of ratios within the tolerances, or $1/n\sum(y_i/x_i)$, where $[a \le (y_i/x_i) \le b]$

(4) The quotient of the sum of the y_i's to the sum of the x_i's within the tolerances, or $\sum y_i / \sum x_i$, where $[a \le (y_i/x_i) \le b]$

(The tolerances are: a=q25-k*H and b=q75+k*H).

Thompson and Sigman (1996) discuss the following plausible model for imputation.

$$Y_i=\beta * X_i+\xi_i, \text{ where } \xi_i | X_i \sim N(0,\sigma^2 X^{\delta})$$

If $\delta=0$ then the best linear unbiased estimator (BLUE) of β is given by $\sum x_i y_i / \sum x_i^2$, the estimator in (1).
If $\delta=1$, the BLUE of β is $\sum y_i / \sum x_i$, the estimator in (4).
If $\delta=2$, the BLUE of β is $1/n\sum(y_i/x_i)$, the estimator in (3).

They note that the estimator given in (2) is often preferred when the imputation goal is to create a consistent distribution of micro data. If the imputation is designed to obtain the correct macro data, however, the estimators (3) and (4) are preferable.

Given a choice of one of the above four types of regression estimators, the SPEER edit system selects as the regression coefficient β_{ij} where the subscript j is determined from $V_j * L_{ij} = \text{Max}\ (V_1 * L_{i1},...,V_t * L_{it})$. If the regression imputation value for $V_i=\beta_{ij} * V_j$ lies within the imputation range for V_i, then this value is imputed for V_i. If the regression imputation value for V_i lies outside the imputation range, the SPEER edit system next selects as the regression coefficient β_{ik} where the subscript k is determined from $V_k * U_{ik} = \text{Min}\ (V_1 * U_{i1},...,V_t * U_{it})$. If the regression imputation value for $V_i=\beta_{ik} * V_k$

lies within the imputation for V_i, then this value is imputed for V_i. If the regression imputation value ($V_i=\beta_{ik}*V_k$) for V_i lies outside the imputation range, then the default imputation scheme is used. This default imputation scheme examines the width of the imputation range. If the range is not too narrow (range≥1 and 1.02*blow≤bup and 0.98*bup≥blow, where blow and bup are the lower and upper bounds of the imputation range), a high/low imputation method is used. If the regression imputation value was less than the lower bound of the imputation range, a value slightly above the lower limit of the range is imputed (blow+max(1.0,0.02*blow)). If the regression imputation value was greater than the upper bound of the range, a value slightly below the upper bound of the range is imputed (bup-max(1.0,0.02*bup)). If it is determined that the imputation range is too narrow, the midpoint of the range is designated as the imputed value.

Additionally, most of the hog variable values (all but the feeder pig variables -- feeder pigs purchased, average price per head, and average weight per head) were rounded to the nearest integer after each iteration.

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 (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.

4. RESULTS FOR QUARTERLY HOG SURVEY

Aggregate statistics from the SPEER edit system are compared with those from the current Blaise/IDAS editing system, which is being treated as "truth." Since a stratified simple random sample of hog operations was selected, each data record corresponding to a hog operation in stratum h was weighted by

$$W_h = N_h/n_{hu}$$

where N_h is the population of hog operations in stratum h, and n_{hu} is the number of usable hog operations in stratum h.

The September 1996 Iowa Key-Entry III file contained only a subset of the cases: those for which data were not collected using CATI. CATI collected data were not included in this study because the data were edited at the time of data capture. The subset of records used in this study contained disproportionately larger hog operations, as can be seen from table 6.

Table 6. Population and Sample Size Counts by Stratum

Stratum	Population N_h	Sample selected n_{hu}	Sample collected on paper
80	4398	91	5
82	9283	366	53
84	7707	549	148
86	2922	419	397
88	950	314	300
92	161	121	117
98	25	25	25

Table 7. Comparison of Expanded Totals for Paper Collected Data Only

Variable	SPEER Edits	Current Procedures	Percentage Difference
Feeder Pigs Purchased	224,372	180,547	24.27
Sows Farrowed 3 mo. Ago	113,029	113,725	-0.61
Sows Farrowed 2 mo. Ago	101,125	101,118	0.01
Sows Farrowed 1 mo. Ago	114,659	114,645	0.01
Sows & Gilts for Breeding	642,985	659,582	-2.52
Sows expected to farrow in next 3 mo.	341,302	341,020	0.08
Sows expected to farrow in 4-6 mo.	315,116	315,448	-0.11
Sows farrowed the last 3 mo.	328,812	329,488	-0.21
Feeder Pig Price	19,616	17,144	14.42
Feeder Pig Lb.	21,538	19,043	13.10
Pig Crop from last 3 mo.	2,547,292	2,557,767	-0.41
Pig Crop from last mo.	937,090	940,284	-0.34
Total Hogs & Pigs	8,109,731	7,107,931	14.09
Market Hogs & Pigs under 60 LBS	1,986,744	1,989,743	-0.15
Market Hogs & Pigs 60-119 LBS	1,671,627	1,675,679	-0.24
Market Hogs & Pigs 120-179 LBS	1,432,136	1,434,399	-0.16
Market Hogs & Pigs 180+ LBS	2,343,304	1,320,620	77.44
Boars & Young Males for Breeding	32,935	27,907	18.02

Table 7. Comparison of Expanded Totals for Paper Collected Data Only (continued)

Variable	SPEER Edits	Current Procedures	Percentage Difference
Pig crop on hand from 3 mo. ago	821,511	822,245	-0.09
Pig crop on hand from 2 mo. ago	799,020	795,239	0.48
Pigs sold or slaughtered from crop 3 mo. ago	135,228	139,037	-2.74
Pigs sold or slaughtered from crop 2 mo. ago	77,576	75,593	2.62
Pigs sold or slaughtered from last mo. crop	48,614	46,631	4.25

Notice that the percentage of samples collected on paper is less than 27 percent in the lowest three strata. By contrast, at least 94 percent of the samples in the remaining strata were collected on paper. The following table shows the expanded values for both systems, and the difference expressed as a percentage of the value from the current system. The expanded totals only include the data for paper collected forms, and do not represent state level indications.

There were 1155 (subtract level) records in the Key-Entry III file. Seventy-one records were excluded in the following tables. Of these, thirteen (1.1%) records required manual editing due to the restriction of simple balance edits in the SPEER edit system; one record still failed edits after six iterations; and fifty seven records were excluded because they were either not considered usable in the final survey data file or because of adjustments made to compensate for frame duplication. The output was obtained using the inputs in Appendix B.

Of the 23 variables, 13 had absolute expanded differences of less than 1 percent. However, six variables (feeder pigs purchased, feeder pig price, feeder pig lb., total hogs & pigs, market hogs & pigs 180lb +, and boars & young males for breeding) had absolute expanded differences exceeding 10 percent.

The large discrepancy between the "total hogs" and "market hogs over 180 pounds" was the result of a single record with a key entry error for the market hogs over 180 pounds. The key entry error duplicated the leading number so that about 1 million too many hogs were keyed in the "180 lbs or over category." The SPEER edit system changed the value of the total hogs variable rather than changing the value of the market hogs over 180 pounds variable. A balance edit was specified where the sum of the market and breeding hog variable values was equal to the total inventory hog variable value. Whenever the sum of the component variable values exceeded the summed variable value for this balance edit, the SPEER edit system was programmed to replace the summed variable value with the sum of the component variable values. Without this record, the absolute expanded difference would have been about 0.10 percent for total hogs and about 1.05 percent for market hogs over 180 lbs.

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

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 price per head to average weight per head exceeded the associated upper ratio edit bound. A higher reliability-weight (2) was assigned to the average price per head variable so that the value of the average weight per head variable would be deleted and imputed. By imputing the value for the average weight per head variable, the resulting values for the average price per head and average weight per head variables were closer to the values for these two variables in other records. The SPEER edit system increased the value of the average weight per head using the relationship between the average weight per head and the average price per head. Thus, the difference in the price per pound in the two systems was only 1.16 percent.

The large percentage difference for boars was the result of SPEER changing the boar inventory for five records. There were no changes made to the boar inventories in the current system. For two of the five records, this value was changed to satisfy the failed balance edit: the sum of market hogs plus the sum of breeding hogs equals the total hog inventory. The balance edit failed since the sum of the market hogs and breeding hogs was less than the total hog inventory. The boar 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 and gilts on hand and the sows and gilts expected to farrow failed. The failed induced edit(s) was derived from the failed ratio edit(s) and the above balance edit.

The moderate percentage differences between the two systems for the pigs sold or slaughtered variables were partially attributable to changes made in the current system. For these records, no edits were failed in the SPEER edit system and hence no changes were made.

The above results show expanded differences in post-edit values between the two systems for survey variables. These results, however, provide no information on the amount of editing performed by the two systems.

Table 8 shows the frequency of records that 1) were not changed in either the SPEER edit system or in the current editing system, 2) not changed in the SPEER edit system but changed in the current editing system, 3) changed in the SPEER edit system but not changed in the current editing system, 4) and changed in the SPEER edit system and changed in the current editing system.

Table 8. Comparison 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			
Feeder Pigs Purchased	989	77	18	0	0
Sows Farrowed 3 mo. Ago	582	495	0	7	0
Sows Farrowed 2 mo. Ago	624	456	0	4	0
Sows Farrowed 1 mo. Ago	590	488	0	5	1

Table 8. Comparison of Same Record Changes (continued)

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			
Sows & Gilts for Breeding	508	554	7	15	0
Sows expected to farrow in next 3 mo.	531	547	5	1	0
Sows expected to farrow in 4-6 mo.	568	507	6	3	0
Sows farrowed the last 3 mo.	529	550	0	4	1
Feeder Pig $/Head	990	77	16	0	1
Feeder Pig Lb./Head	990	75	1	1	17
Pig Crop from last 3 mo.	537	543	4	0	0
Pig Crop from last mo.	595	486	3	0	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 180+ LBS	415	654	12	3	0
Boars & Young Males for Breeding	530	549	0	5	0
Pig crop on hand from 3 mo. ago	625	454	4	1	0
Pig crop on hand from 2 mo. ago	643	436	3	2	0
Pigs sold or slaughtered from crop 3 mo. ago	1015	66	2	1	0
Pigs sold or slaughtered from crop 2 mo. ago	1046	37	1	0	0
Pigs sold or slaughtered from last mo. crop	1059	24	1	0	0
Sum of all variables	14964	9773	107	62	26

The entries in Table 8 reveal that the two systems usually did not make changes to the same record variable values, except for the feeder pig average weight per head. The current 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 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 (see Table 8). Note that the current system changed some

values for the pigs sold or slaughtered variables when the SPEER edit system did not. The few changes made to these values resulted in expanded differences ranging from 2.5 to 4.25 percent in Table 8.

The following table shows the number of records for each survey variable that showed an increase in value, a decrease in value, and no change. If changes are consistently positive or negative, this may indicate that either the editing process is biased, or that there are measurement errors associated with the questionnaire: the words in the question, the structure of the question, and the order or context of questions.

Table 9. Comparison of the Direction of Changes

Variable	SPEER Edit System			Current System		
	Increase	Decrease	No Change	Increase	Decrease	No Change
Feeder Pigs Purchased	0	0	1084	0	18	1066
Sows Farrowed 3 mo. Ago	1	6	1077	0	0	1084
Sows Farrowed 2 mo. Ago	2	2	1080	0	0	1084
Sows Farrowed 1 mo. Ago	2	4	1078	0	1	1083
Sows & Gilts for Breeding	3	12	1069	5	2	1077
Sows expected to farrow in next 3 mo.	1	0	1083	2	3	1079
Sows expected to farrow in 4-6 mo.	1	2	1081	2	4	1078
Sows farrowed the last 3 mo.	1	4	1079	0	1	1083
Feeder Pig $/Head	1	0	1083	0	17	1067
Feeder Pig Lb./Head	16	2	1066	0	18	1066
Pig Crop from last 3 mo.	0	0	1084	3	1	1080
Pig Crop from last mo.	0	0	1084	2	1	1081
Total Hogs & Pigs	4	5	1075	8	4	1072
Market Hogs & Pigs under 60 LBS	1	1	1082	5	2	1077
Market Hogs & Pigs 60-119 LBS	1	0	1083	5	1	1078
Market Hogs & Pigs 120-179 LBS	2	2	1080	4	1	1079
Market Hogs & Pigs 180+ LBS	3	0	1081	5	7	1072
Boars & Young Males for Breeding	5	0	1079	0	0	1084
Pig crop on hand from 3 mo. ago	1	0	1083	3	1	1080

Table 9. Comparison of the Direction of Changes (continued)

Variable	SPEER Edit System			Current System		
	Increase	Decrease	No Change	Increase	Decrease	No Change
Pig crop on hand from 2 mo. ago	2	0	1082	2	1	1081
Pigs sold or slaughtered from crop 3 mo. ago	0	1	1083	1	1	1082
Pigs sold or slaughtered from crop 2 mo. ago	0	0	1084	0	1	1083
Pigs sold or slaughtered from last mo. crop	0	0	1084	0	1	1083

From the entries in Table 9, it is clearly seen that the large majority of records for both systems had no change made to the variable values. Notice that since all changes made to the feeder pig variable values by the current system are negative (the values were zeroed out), there may be problems associated with the questionnaire.

Also noteworthy is the large number of changes (mostly negative) made by the SPEER edit system to the values of the sows and gilts for breeding variable. This was the result of the edit limits for the ratio of sows and gilts for breeding to either the sows expected to farrow in the next three months or in four to six months being too restrictive. Usually, the ratio edits involving these variables failed because the ratio value was greater than the upper edit bound (The upper ratio edit bounds were 4.3 and 4.8, respectively). The SPEER edit system, in adjusting the expected farrowing rates, changed the value of the sows and gilts for breeding variable.

5. RESULTS FOR QUARTERLY AGRICULTURAL SURVEY

Since the edits for the hog data contained several balance edits that violated the simple balance edit restriction, it was decided to examine data from the December Agricultural Survey, a data set thought to require fewer balance edits. However, when formulating the edits for the SPEER edit system, two balance edits requiring the same variable were needed, thus violating the simple balance edit restriction. In addition, a balance edit of the form $V_1+V_2-V_3=V_4$ was needed. Balance edits are specified in the SPEER edit system as $\sum_i V_i=V_s$. To formulate the above balance edit, two user-defined variables need to be created: $V_5=V_1+V_2$ and $V_6=V_3+V_4$ In addition, the ratio edit $1 \le V_5/V_6 \le 1$ needs to be specified.

Many of the edits (more so than with the hog data set) were of the form $L \le V_i \le U$, where the value of variable V_i is required to be greater than or equal to a lower limit L and less than or equal to an upper limit U. Since the edits in the SPEER edit system must be formulated either as ratio or balance edits, a dummy variable, V_j, with the value of one was created so that the edit could be formulated as a ratio edit of the form $L_{ij} \le V_i/V_j \le U_{ij}$. This led to the generation of a rather large number of implicit ratio edits. Sixty-nine explicit ratio edits were specified resulting in 143 ratio edits in the complete set of ratio edits.

The examination of this data set provided more information on the usefulness of using several iterations to correct data records. Although the ratio edit bounds generated by the resistant fences may have been too

restrictive for some ratio edits resulting in more failed edits, the SPEER edit system failed to correct over seven percent of the data records even after six iterations. One reason that the SPEER edit system failed to correct some records was that all of the connected data were missing.

To understand how this could happen, a particular record is considered. In order to formulate the balance edit $V_{27}+V_{28}-V_{29}=V_{32}$ ("land owned" + "land rented from" - "land rented to" = "total land"), the following edits were specified $V_{30}=V_{27}+V_{28}$, $V_{31}=V_{29}+V_{32}$, and $1 \le V_{30}/V_{31} \le 1$. Variables V_{31} and V_{32} were identified in the error localization problem to have their values deleted. Since the SPEER edit system imputes variable values in ascending order, the value for variable V_{31} is imputed prior to that of variable V_{32}. The edits involving variable V_{31} are $V_{31}=V_{29}+V_{32}$ and $1 \le V_{30}/V_{31} \le 1$. The value for variable V_{31} cannot be imputed using the balance edit since the value for variable V_{32} has been deleted in the solution to the error localization problem. Nor can the value for variable V_{31} be imputed from the ratio edit because the reported value for variable V_{30} was zero (so, all connected data are missing). Thus, the value for variable V_{30} will not be changed leaving the ratio edit unsatisfied. If, however, ratio edits were allowed to fail regardless of the values of the variables involved, many ratio edits with a positive lower edit bound would fail if the value of the variable in the numerator had a zero value. This would likely lead to many variable values that had reported zero values being replaced with positive imputed values -- an undesirable event. Many variable values in NASS surveys are valid zeroes. This is not the case for surveys at the Bureau of the Census for which the SPEER edit system is used, where the variable values are generally positive.

Two other problems with SPEER that led to the failure to correct data records after six iterations are now mentioned. The first problem is that the imputation ranges for some variables were negative because the variable values used to construct the ranges were imputed using a balance edit. These imputed values were the same as the reported values. Thus, the originally failed ratio edits would still fail. The second problem arises when the edit $1 \le V_{30}/V_{31} \le 1$ is satisfied and the value for variable V_{31} is identified to be deleted in the error localization problem. Since the only variable in the connected set involving variable V_{31} is variable V_{30}, the imputation range for variable V_{31} consists of a single value, namely, the value of variable V_{30}. This problem is similar to the first problem in that the imputed value is the reported value which caused edits to fail in the first place. Since the imputed values are the same as the reported values, the SPEER edit system fails to correct the data record iteration after iteration.

6. CONCLUSIONS AND RECOMMENDATIONS

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

1) 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.

2) Commodity data editing and imputation are performed by the system. This minimizes the need for manually reviewing and correcting the data records. Statisticians would only need to review the 1 to 7 percent of records that the SPEER edit system could not handle and the records where the SPEER edit system

made extremely large changes (which should then be manually reviewed).

3) The system provides an audit trail. That is, it keeps track of the changes made and the reasons for making the changes. This can be useful 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.

4) 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 133 Mhz Pentium computer.

5) The results of the system are repeatable in time and space. That is, the results of data records run through the SPEER edit system will be the same regardless of when they were run through the system and the location of the system.

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

1) 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., completion code presence/absence editing) and post-SPEER edits (corrections for records the SPEER edit system could not handle) must be performed outside of the system.

2) Undesirably large changes may be made to variable values. One record accounted for nearly the entire 14 percent difference in expanded total hogs and pigs for the two systems. The system could be modified to output those records for which this occurs for manual review (as the word referrals suggests in the acronym SPEER), but this increases the number of records not handled by the system. Moreover, editors may lose confidence in the SPEER edit system's procedures when they see only these extreme values.

3) Because not all commodity edits can be handled by the SPEER edit system, the logistics of implementation are difficult. The use of multiple systems to incorporate all edits results in data records being cycled from system to system leading to potential frustrations for statisticians.

4) The SPEER edit system is a black box to the editor. It is not always easy to understand the actions of the system, especially with the generation of implied edits to identify variable values to delete and impute. Setting up the explicit ratio edits and understanding the implied edits is not the "natural" way of specifying edits.

Therefore, the following recommendations are made:

1) NASS should not use the current SPEER edit system for editing survey data because the specification of edits is quite restricted for NASS type data. Only a very limited subset of balance edits are accommodated. And even for this subset, the error localization problem may not be properly solved since a complete set of edits is not generated. Thus, the SPEER edit system is only useful 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 SPEER edit system.

2) Investigate the potential of periodically using resistant fences to specify edit bounds for current computer edits (Blaise or SPS). This might help "standardize" the limit specification.

3) Investigate other edit systems which use historical profile data, such as the livestock slaughter or Ag-Complex as potential tools. This type of edit system would be restricted to large operations and agribusinesses. However, the concurrent development of the data warehouse should serve as an invaluable resource for the development of a profile based editing system.

Investigation of GEIS, a more generalized approach to the SPEER edit system, is not recommended because resources are not available for modifying the system for a non-ORACLE environment.

REFERENCES

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

Fellegi, I.P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical System*, 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. (1982), "Examples Illustrating the Need for Implied Edits for Continuous Data," unpublished documentation, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

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

Greenberg, B.G. (1986), "The Use of Implied Edits and Set Covering in Automated Data Editing," Statistical Research Division report RR-86/02, 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.

Groves, R.M. (1989), *Survey Errors and Survey Costs*. NY: Wiley.

Hanuschak, G. 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.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (Eds.) (1983), *Understanding Robust and Exploratory Data Analysis*. NY: Wiley.

Hoaglin, D.C., Iglewicz, B., and Tukey, J.W. (1986), "Performance of Some Resistant Rules for Outlier Labeling," *Journal of the American Statistical Association*, 81, 991-999.

Imel, J., and Ramos, M. (1992), "Results of Follow-up Study Comparing the Performance of the Ag-Complex Edit and Imputation System and the Ag-SPEER Edit and Imputation System," Bureau of the Census: Agriculture Division. Internal documentation.

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.

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

Pierzchala, M. (1990), "A Review of Three Editing and Imputation Systems," National Agricultural Statistics Service, USDA, Washington D.C.

SAS Institute (1990a), *SAS/GRAPH Software*, Version 6, First Edition, Vol. 1-2. Cary, NC: SAS Institute.

SAS Institute (1990b), *SAS/OR User's Guide*, Version 6, First Edition. Cary, NC: SAS Institute.

SAS Institute (1991), *Guide to Macro Processing*, Version 6, Second Edition. Cary, NC: SAS Institute.

SAS Institute (1992), *Procedures Guide*, Version 6, Third Edition. Cary, NC: SAS Institute.

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.*

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

APPENDIX A: EXAMPLE OF IMPROPERLY SOLVING THE ERROR LOCALIZATION PROBLEM

Below is a simple example illustrating the case when the minimal set of variable values identified in the solution to the error localization is not enough.

The explicitly specified ratio edits are:

2.5000000 < EMP1 / APR2 < 3.5000000
1.5000000 < EMP1 / QPR3 < 3.5000000
2.0000000 < EMP1 / FBR4 < 4.0000000

The specified balance edit is:

EMP1 = APR2 + QPR3 + FBR4

SPEER Output:

Record # 1

Failed edits:
2.5000000 < EMP1 / APR2 < 3.5000000
1.5000000 < EMP1 / QPR3 < 3.5000000
0.5714286 < APR2 / FBR4 < 1.6000000

At least one balance equation is not satisfied.
At least one induced edit is not satisfied.

Deleted fields: 1. EMP1 2. APR2

**** EMP1 imputation range unacceptable ****
Lo= 35.58800 Up= 28.04000

Imputation range for APR2: Lo = 10.1680 Up = 11.2160
APR2 imputed using hi - lo default: hi option

Bounds computed using reported and imputed values

Fields	Revised	Reported	Lower	Upper
EMP1	-1.000	21.460	35.588	28.040
APR2	10.216	17.370	10.168	-11.945
QPR3	18.410	18.410	8.314	-17.601
FBR4	7.010	7.010	6.904	-19.124

The minimal set identified is {EMP1, APR2} since all of the failed edits - ratio, balance, and induced - involve either EMP1 or APR2. There is no possible way to satisfy all edits with this minimal deletion set. If however, three variables are deleted, say, {EMP1, APR2, QPR3}, then the variable values EMP1=26.01, APR2=9, QPR3=10, and FBR4=7.01 satisfy all edits. (Ignore the final variable bounds as they are meaningless since the value for the variable EMP1 is missing).

If the SPEER edit system had generated the nonsimple induced edits, the error localization problem would have been correctly solved. In particular, the failure of the nonsimple induced edit derived below would have led to deleting and imputing the value of either the variable QPR3 or FBR4. (It is noted, however, that the minimal number of fields to change so that all edits are satisfied is two).

Rewrite the third explicit ratio as:
$2FBR4 \leq EMP1 \leq 4FBR4$

Substituting APR2+QPR3+FBR4 for EMP1 from the balance edit yields:
$2FBR4-APR2-QPR3 \leq FBR4 \leq 4FBR4-APR2-QPR3$

Now, substituting for APR2 using $0.571FBR4 \leq APR2 \leq 1.6FBR4$ yields the following nonsimple induced edit:
$2FBR4-1.6FBR4-QPR3 \leq FBR4 \leq 4FBR4-0.571FBR4-QPR3$
which is equivalent to
$0.4FBR4-QPR3 \leq FBR4 \leq 3.429FBR4-QPR3$.

Substituting the originally reported variable values into the nonsimple induced edit yields:
$(0.4)*(7.01)-18.41 \leq 7.01 \leq (3.429)*(7.01)-18.41$
which is equivalent to
$-15.61 \leq 7.01 \leq 5.63$.

Clearly, the nonsimple induced edit fails, and as such, either the value of variable QPR3 or FBR4 must be deleted and subsequently imputed so that the edit would be satisfied.

APPENDIX B: RUNNING THE SPEER EDIT SYSTEM USING QUARTERLY HOG SURVEY DATA

Detailed instructions on how to run the SPEER edit system are provided using data from the September 1996 Iowa hog survey. The first step, which is to identify the edits, was greatly facilitated by the HEAT team's report entitled, "Report of the Hog Editing and Analysis Team" (see Anderson et al., 1996). Each edit in this report is classified as a warning or a critical edit. A warning edit in the current Blaise/IDAS editing system may be suppressed by the editor, whereas the more severe, critical edit generally requires the user to enter or modify survey data. Since warning edits are usually suppressed, only the critical edits were used in the SPEER edit system. These critical edits need to be reformulated into ratio and balance edits as required by the SPEER edit system. Each (numbered) critical edit recommended by the HEAT team is listed below followed by its reformulation in the SPEER edit system.

501: Total hogs cannot be positive when the sum of the classes is zero.
520: Total hogs must equal the sum of the classes.
SPEER: Balance edit

lhgund60 + lhgto119 + lhgto179 + lhgov180 + lhoggilt + lhogboar = lhogtotl

505: Sows and gilts expected to farrow in each of the next two quarters must come from sows and gilts on hand.
SPEER: Ratio edits lhoggilt / lhgexp13; lhoggilt / lhgexp46

508: Recorded pigs on hand and pigs sold or slaughtered must have been from litters previously farrowed.
SPEER: Ratio edits lhgpgsld / lhgfar13; lhpgsld1 / lhogfar1; lhpgsld2 / lhogfar2;
lhpgsld3 / lhogfar3

where lhgpglsd, lhpgsld1, lhpgsld2, and lhpgsld3 are user-defined as follows:

lhgpglsd = lhgpig13 + lhgsld13
lhpgsld1 = lhogpig1 + lhogsld1
lhpgsld2 = lhogpig2 + lhogsld2
lhpgsld3 = lhogpig3 + lhogsld3

510: Pigs one month of age or younger on hand must fall into either market under 60 lbs, boars, or sows and gilts.
SPEER: Ratio Edit lhogpig3 / lhg60brd, where lhg60brd is user defined as follows:

lhg60brd = lhgund60 + lhogboar + lhoggilt

511: Pigs on hand from previous three months litters are normally in the lighter three market weight groups or in the breeding classes.
SPEER: Ratio edit lhgpig13 / lhg180br, where lhg180br is user-defined as follows:

lhg180br = lhgund60 + lhgto119 + lhgto179 + lhogboar + lhoggilt

518: If the pigs recorded from previous farrowings are "truly" on hand then they must be recorded in inventory.

SPEER: Ratio edit lhgpig13 / lhogtotl

536: All items must be present: number purchased, price per head, and weight per head.
SPEER: Ratio Edits lhgfdpur / lhgfdcst ; lhgfdpur / lhgfdlbs

537: Pigs with an average weight per head outside limits should not be considered feeder pigs.
SPEER: Ratio edit lhgfdlbs/dumone, where dumone is user-defined as: dumone = 1

538: Derived price per pound is less than $.20 or greater than $2.00.
SPEER: Ratio edit lhgfdcst / lhgfdlbs

The critical edit 531: Hog completion code is 2 and previous quarter inventory or control data is greater than 1,000 is not easily handled in the SPEER edit system. This type of edit should be handled as a pre-edit and can easily be done using a program written in SAS.

From the above edits, the resulting balance edits are:

lhogfar1 + lhogfar2 + lhogfar3 = lhgfar13
lhgund60 + lhgto119 + lhgto179 + lhgov180 + lhoggilt + lhogboar = lhogtotl
lhogpig1 + lhogsld1 = lhpgsld1
lhogpig2 + lhogsld2 = lhpgsld2
lhogpig3 + lhogsld3 = lhpgsld3
lhogpig1 + lhogpig2 + lhogpig3 = lhgpig13
lhogpig1 + lhogpig2 + lhogpig3 + lhogsld1 + lhogsld2 + lhogsld3 = lhgpgsld
lhoggilt + lhogboar + lhgund60 = lhg60brd
lhoggilt + lhogboar + lhgund60 + lhgto119 + lhgto179 = lhg180br

Since the SPEER edit system can accommodate only simple balance edits (see Section 3.1.1), the last four balance edits were excluded. Thus, for any data records, if any of the four variables, lhgpig13, lhgpglsd, lhg60brd, lhg180br, are identified in the error localization problem to have their values deleted, the corresponding component variables in the four balance edits will not have their values updated. These data records must be edited by some other means since they are not adequately edited and imputed in the SPEER edit system. (Thirteen such records were identified in Section 4).

Once the edits have been identified, the ratio edit bounds can be calculated using the resistant fences procedure using the SAS program, res_bnd.sas. Thompson and Sigman (1996) provide instructions on how to run this program. They identify the following general steps:

1. Prepare Input SAS Data Set
2. Modify res_bnd.sas for Run
   Modify Global Macro Variables
   Specify Ratio Tests
   (Optional) Specify Value of Classification Variable
3. Save the Program
4. Submit the Program
5. (Optional) Examine Outputs

A SAS data set was created using final Iowa hog survey data files from June 95, September 95, December 95, and March 96. After concatenating these files, the resulting file was subsetted to exclude refusals and inaccessables, keeping only variables of interest. Also, the value of the dummy variable, dumone, was set to 1. Finally, if any component variable values that are added together to create a user-defined variable value are missing, these values are set equal to zero so that the user-defined variable value would not be assigned a missing value. As an example, consider the user-defined variable, lhg60brd, which is defined as the sum of the variables, lhoggilt, lhogboar, and lhgund60. If any one of these latter three variables had a missing value, then the value of the sum of these variables, equal to the value of the variable lhg60brd, would also be missing. For those ratios involving variables whose values are asked on a monthly basis (e.g., lhpgsld1/lhogfar1), the June 96, final Iowa hog survey data file was used to create the ratio edit bounds. This was the first month in which data were asked monthly.

The global macro variables were set to the following values, taken from the res_bnd.sas program. Refer to Thompson and Sigman (1996) for more details.

```
***** PREPARATION FOR PROGRAM ** GLOBAL VARIABLES ******;

*Specify Input and Output Locations and Data Set Names;

%global inlib; * the location of the input SAS Data set;
%global filename; * the name of the input SAS Data set;
%global outlib; * the location of the output SAS Data set (summary file);
%global rpf_loc; * the location of the ratio edit parameter (ascii) file;
%global outfile; * the name of the output SAS Data set and ratio edit pf;
%global iset_loc; * the location of the SAS/INSIGHT Data set of ratios;
%global graphout; * the location of the output SAS/GRAPH .gsf file;
%global version; * the version of SAS being used;

%let inlib=f:\users\todato\;
%let filename=ia96inp;
%let outlib=f:\users\todato\edit\;
%let rpf_loc=f:\users\todato\edit\;
%let outfile=editfile;
%let graphout=f:\users\todato\edit\;
%let iset_loc=f:\users\todato\edit\;
%let version=610;
```

```
*Specify Program Options for Outputs;

%let keep_ds=y; * Create Permanent SAS Data set (Y/N);
%let allrbar=n; * Include All Three Estimates of RBAR in Permanent SAS Data set (Y/N);
%let sumstat=n; * Produce Printout of Summary Statistics (Y/N);
%let univstat=y; * Produce Printout of Univariate Statistics for Each Ratio (Y/N);
%let graphbnd=n; * Produce SAS/Graph of Reported Data Versus Bounds (Y/N);
%let plot_rat=n; * Produce ASCII Plot of Reported Data Versus Bounds (Y/N);
%let insgtset=n; * Create analysis Data set of ratios (Y/N);

* Specify Names of Unique Identifiers on Input SAS Data set;

%global c_name; * Name of classification variable;
 * If c_name is blank, then entire file is used for bounds;
%global e_ident; * Name of unique establishment identifier;
 * This can be blank;
%let c_name=mpreshog;
%let e_ident=id;

* Specify How To Process Input SAS Data set;

%let spec_cls=y; * Specify a value of classification variable (Y/N);
 * if Y, then go to macro begin;
 * The default is to automatically create bounds for
 * each value of classification variable in the input file (N);

* Specify Approach/Parameters for Ratio Edit Tolerances and Industry Avg;

%global k_val; * The Value of K Used for Resistant Fences;

%let k_val=3;
%let sym_dist=n; * Symmetrize Distribution (Y/N);
%let rbartype=blue; * Type of Industry Average (SUM,AVG,MED,BLUE);
%let neg_sub=z; * Value substituted for negative lower bound (Z/F);
```

The ratio edits are specified using SAS inputs statements. The numerator variable is the first parameter specified followed by denominator variable. The SAS code is shown below.

```
*Hog Survey ratios;
%inputs(lhoggilt,lhgexp13); /* r_505 Blaise; */
%inputs(lhoggilt,lhgexp46); /* r_505 Blaise; */
%inputs(lhgpgsld,lhgfar13); /* r_508 Blaise; */
%inputs(lhpgsld1,lhogfar1); /* r_508 Blaise; */
%inputs(lhpgsld2,lhogfar2); /* r_508 Blaise; */
%inputs(lhpgsld3,lhogfar3); /* r_508 Blaise; */
%inputs(lhogpig3,lhg60brd); /* r_510 Blaise; */
```

%inputs(lhgpig13,lhg180br); / r_511 Blaise; */*
%inputs(lhgpig13,lhogtotl); / r_518 Blaise; */*
%inputs(lhgfdpur,lhgfdcst); / r_536 Blaise; */*
%inputs(lhgfdpur,lhgfdlbs); / r_536 Blaise; */*
%inputs(lhgfdlbs,dumone); / r_537 Blaise; */*
%inputs(lhgfdcst,lhgfdlbs); / r_538 Blaise; */*

The option to specify the classification variable allows for the calculation of ratio edit bounds for each unique value of the classification variable.

The location and the name of the output file containing the ratio edit bounds are assigned when specifying the values of the global macro variables. The location of the output file given above is: *%let outlib = f:\users\todato\edit\;*. The output file above is named using the following statement: *%let outfile = editfile*. The contents of this file include five variables (columns):

1. Class: the value of the classification variable
2. Ratio: the mnemonic ratio of two variables
3. Lower edit bound
4. Upper edit bound
5. Industry Average: An average measure of the ratio values

Sample output of this file is shown below.

0	lhoggilt/lhgexp13	1.0000000	4.3000000	1.7948311
0	lhoggilt/lhgexp46	1.0000000	4.7777778	1.8193021
0	lhgpgsld/lhgfar13	3.0000000	13.5000000	8.9479801
0	lhgpig13/lhg180br	0.0000000	1.3274376	0.4688133
0	lhgpig13/lhogtotl	0.0000000	1.0000000	0.3575552
0	lhgfdpur/lhgfdcst	0.0100000	150.0000000	5.4672522
0	lhgfdpur/lhgfdlbs	0.0100000	150.0000000	3.5526131
0	lhgfdlbs/dumone	0.0000000	120.0000000	50.9568733
0	lhgfdcst/lhgfdlbs	0.0000000	1.6689866	0.6848262
0	lhpgsld1/lhogfar1	4.7272727	12.3636363	9.2955536
0	lhpgsld2/lhogfar2	4.7272727	12.3636363	9.4077743
0	lhpgsld3/lhogfar3	4.2500000	13.0000000	9.7309520
0	lhogpig3/lhg60brd	0.0000000	1.0000000	0.3609643

Several comments are now made concerning the above output.

The res_bnd.sas program was intended to calculate ratio edit bounds for positive data. If the value of the numerator variable was zero or missing, or the value of the denominator was zero or missing, for a particular record, then the record was excluded in the calculation of the ratio edit bounds.

Several of the ratio edit bounds were modified to include subject knowledge about the ratios. The lower ratio edit bounds of the ratios lhoggilt/lhgexp13 and lhoggilt/lhgexp46 were assigned values of one. This was done since the number of sows and gilts must be at least as large as the number of sows and gilts expected to farrow in the future. The upper ratio edit bound of the ratio lhogpig3/lhg60brd was assigned a value of one since the pig crop kept on hand from the previous month's pig crop cannot exceed the sum of the market hogs under 60 pounds and the hogs kept for breeding. A value of one was used as the upper ratio edit bound for the ratio lhgpig13/lhogtotl since the pig crop on hand from the previous quarter must be accounted for in the total hog inventory. The lower ratio edit bounds for the ratio edits lhgfdpur/lhgfdcst and lhgfdpur/lhgfdlbs were assigned values of 0.01. The reason for this assignment was to satisfy the Heat Team's edit number 536:

536: All items must be present: number purchased, price per head, and weight per head.
SPEER: Ratio Edits lhgfdpur / lhgfdcst ; lhgfdpur / lhgfdlbs

and to conform to the programming of this edit in Blaise. This edit was programmed into Blaise as follows:

If lhgfdcst > 0 then lhgfdpur >0 ; If lhgfdlbs > 0 then lhgfdpur > 0.

The only way to formulate the first (second) IF-THEN statement into a ratio edit is to assign a small positive value to the lower bound of the ratio lhgfdpur/lhgfdcst (lhgfdpur/lhgfdlbs). Thus, if the value of the variable lhgfdcst (lhgfdlbs) is greater than zero, then value of the variable lhgfdpur must be greater than zero in order for the ratio edit to be satisfied. The upper ratio edit bounds for these two ratio edits were assigned values of 150.

An auxiliary SAS program was written to modify the above output for input into the FORTRAN program gb3.for which calculates the implicit ratio edits. The output of the SAS program is written to the file ratios.exp and is shown below. The first column is value for the classification variable. The second column is a list of ascending numbers from one to the number of ratio edits. These numbers are assigned to the ratio edit for identification purposes. The numbers in the third and fourth columns correspond to the numerator variable and the denominator variable, respectively. The fifth and sixth columns contain the lower and upper ratio edit bounds, respectively. The seventh column contains an average measure of the ratio values. Finally, the eighth column is the mnemonic ratio of the two variables.

0	1	1	11	0.01000	150.000	5.46725	lhgfdpur/lhgfdcst
0	2	1	12	0.01000	150.000	3.55261	lhgfdpur/lhgfdlbs
0	3	2	17	4.72727	12.363	9.29556	lhpgsld1/lhogfar1
0	4	3	18	4.72727	12.363	9.40778	lhpgsld2/lhogfar2
0	5	4	19	4.25000	13.000	9.73095	lhpgsld3/lhogfar3
0	6	5	8	1.00000	4.300	1.79483	lhoggilt/lhgexp13
0	7	5	9	1.00000	4.778	1.81930	lhoggilt/lhgexp46
0	8	10	13	3.00000	13.500	8.94798	lhgpgsld/lhgfar13
0	9	11	12	0.00000	1.669	0.68483	lhgfdcst/lhgfdlbs
0	10	12	20	0.00000	120.000	50.95687	lhgfdlbs/dumone
0	11	14	6	0.00000	1.327	0.35803	lhgpig13/lhg180br
0	12	14	16	0.00000	1.000	0.35756	lhgpig13/lhogtotl
0	13	15	7	0.00000	1.000	0.36096	lhogpig3/lhg60brd

The FORTRAN program gb3.for logically derives implied ratio edits, termed implicit ratio edits, from the explicit ratio edits in the ratios.exp file and writes the complete set of ratio edits to the file ratios.imp. This program reads a file named filename.dat containing the following:

RATIOS.EXP
RATIOS.IMP
STATUS
(I4,4X,2I4,2F16,7)

The first two lines contain the names of the files ratios.exp and ratios.imp which are the input and output files, respectively, containing ratio edits. The third line contains the name of the file status to which any encountered problems are written. The fourth line is the FORTRAN format that is used to read the contents of the file ratios.exp. Finally, it is noted that the FORTRAN parameter statement located in the gb3.for program,

```
PARAMETER(BFLD=40)
```

is used to specify the maximum number of distinct variables involved in the explicit ratio edits.

The contents of the file ratios.imp is shown below. The first column is the classification variable value. The next column is the mnemonic ratio followed by the variable numbered ratio. Finally, the lower and upper ratio edit bounds are listed. The large upper ratio edit bounds indicate that the upper bound is unlimited.

0	lhgfdpur/lhgfdcst	1,11	0.0100000	150.0000000
0	lhgfdpur/lhgfdlbs	1,12	0.0100000	150.0000000
0	lhgfdpur/dumone	1,20	0.0000000	18000.0000000
0	lhpgsld1/lhogfar1	2,17	4.7272727	12.3636360
0	lhpgsld2/lhogfar2	3,18	4.7272727	12.3636360
0	lhpgsld3/lhogfar3	4,19	4.2500000	13.0000000
0	lhoggilt/lhgexp13	5,8	1.0000000	4.3000000
0	lhoggilt/lhgexp46	5,9	1.0000000	4.7777777
0	lhg180br/lhgpig13	6,14	0.7533311	999999.0000000
0	lhg60brd/lhogpig3	7,15	1.0000000	999999.6870000
0	lhgexp13/lhgexp46	8,9	0.2325581	4.7777777
0	lhgpgsld/lhgfar13	10,13	3.0000000	13.5000000
0	lhgfdcst/lhgfdlbs	11,12	0.0000667	1.6689866
0	lhgfdcst/dumone	11,20	0.0000000	200.2783810
0	lhgfdlbs/dumone	12,20	0.0000000	120.0000000
0	lhgpig13/lhogtotl	14,16	0.0000000	1.0000000

The FORTRAN program, cmpbeta3.for, is used to calculate the ratio regression coefficients which are used in the imputation process. The Key-Entry III data file to be edited is also used as the data from which these coefficients are calculated. A brief digression is now made to discuss the preparation of the Key-Entry III data file. A SAS program reads the data which consists of an item code corresponding to a variable followed by its value and creates a SAS data file. Any user-defined variables (e.g., lhg60brd) are also created by summing the associated component variables. Finally, this SAS data file is compared with the final survey data file to resolve any duplicate records and to identify and fix any anomaly records (e.g., a refusal record with positive data). The SAS data file is then written to an ASCII data file speer.dat containing variable values to be edited and any other informative variables that may be helpful in interpreting the output from the SPEER edit system (i.e., state, id, tract, subtract, mrespons). This data file can now be used as input into the program cmpbeta3.for.

The maximum number of variables and records is specified in cmpbeta3.for via a FORTRAN parameter statement. The following parameter statement was used for the Iowa Key-Entry III file which allows for a maximum of 35 variables and 1300 records: Parameter(bfld=35, brecs=1300). The program cmpbeta3.for reads as input the contents of a file named betaname.dat. This file contains the following five lines.

SPEER.DAT
RATIOS.EXP
BETA.DAT
(I6,30,F14.3)
30

As mentioned above, speer.dat is the Key-Entry III data file. The file ratios.exp contains the explicit ratio edits. The SPEER edit system uses this file as input to determine for which pair of variables a ratio regression coefficient will be calculated. The cmpbeta3.for program generates ratio regression coefficients for all pairs of variables that are involved in the same connected set (Section 3.3). The filename beta.dat is the file to which the ratio regression coefficients are written. The fourth line is the FORTRAN format for reading the speer.dat file. The last line once again specifies the number of variables in the file speer.dat.

As mentioned earlier, the res_bnd.sas program was used to assist in the calculation of the ratio regression coefficients. After examining the beta.dat file, it was noticed that some of the ratio regression coefficients were very small (or large), particularly the coefficients associated with the feeder pig variables. The res_bnd.sas program offered the potential to detect and eliminate outliers in the calculation of these coefficients. However, res_bnd.sas did not calculate an average measure using the least squares method that cmpbeta3.for used. Thus, the least squares option was added to the program res_bnd.sas so that it could be used to calculate ratio regression coefficients with outliers eliminated. After comparing the ratio regression coefficients from cmpbeta3.for and res_bnd.sas, it was determined that the coefficients were similar except for the coefficients associated with the feeder pig variables and the user-defined variable lhg180br. Further analysis identified that a few observations highly influenced the value of these coefficients. These observations had variable values that were keyed in the wrong units. To eliminate the effects of the outliers, the coefficients calculated from cmpbeta3.for were replaced with the coefficients calculated from res_bnd.sas.

The main SPEER program, spr3d.for, reads as input the file named sprname.dat. The contents of sprname.dat are as follows:

SPEER.DAT
SPR.OUT
SUM.OUT
(I6,30F14.3,I4,I11,3I4)
30
RATIOS.IMP
BNAMES.DAT
BETA.DAT
WGTS.DAT

As before, speer.dat is the Key-Entry III file that is to be edited. The file spr.out contains indvidual data record summaries. The file sum.out is a summary file containing information on the frequency of failed ratio edits and the frequency of deleted variable values. The fourth line is a FORTRAN format that is used to read in the file speer.dat. (The I6 symbol corresponds to the classification

variable value. The symbol 30f14.3 indicates that there are thirty variables values that are to be read next. The I4 symbol corresponds to the state FIPS code. The I11 symbol corresponds to the operation identification number. The 3I4 symbol corresponds to the tract, subtract and mrespons variables.) The fifth line indicates that thirty variables are to be read from the files speer.dat and bnames.dat. The sixth line is the name of the file ratios.imp that contains the complete set of ratio edits. The file bnames.dat is listed on the seventh line. This file contains the mnemonic variable names. The file beta.dat contains the ratio regression coefficients. The ninth and final line contains the name of the file wgts.dat. This file contains a weight for each variable indicating the reliability of the value reported for the variable; the higher the weight, the higher the reliability and the less likely the variable value will be deleted. The default variable weights are 1.00. The weight associated with the variable dumone was assigned the value 9999.99. Thus, with such a large value, this variable value will never be deleted. Since this is not a survey variable, its value should never be changed to satisfy a ratio edit involving this variable. Also, the weights for the expected farrowing variables were set to 5.00.

Balance edits may be specified in a file sum.dat which is read as input by the program spr3d.for. The contents of this file are shown below.

```
5
4 3 4 5 11
7 21 22 23 24 6 25 17
3 26 28 18
3 27 29 19
3 16 30 20
```

The number 5 on the first line indicates that five balance edits are to follow. In each of the following five lines, the first number denotes the number of variables involved in each balance edit. The second to the penultimate number correspond to the component variables. The last number corresponds to the summed variable.

Four output files are produced by the spr3d.for program: check out, edit.out, spr.out and sum.out. The output file check.out was added to the program and contains those data records that the SPEER edit system could not properly handle after six iterations. The file edit.out contains the edited speer.dat (Key-Entry III) data file. The file spr.out lists for each data record run through the SPEER edit system the failed ratio edits, whether or not one or more balance edits failed, whether or not one or more induced edits failed, the variable values deleted, an imputation range for each deleted variable, the imputation scheme employed to impute a value, record information, the original variable values and the imputed variable values. Listed below is a record from the file spr.out.

```
Record # 52

Failed edits:
1.0000000 < lhoggilt / lhgexp13 < 4.3000002
1.0000000 < lhoggilt / lhgexp46 < 4.7777777
At least one induced edit is not satisfied.
```

Deleted fields: 23. lhgto179 5. lhoggilt

Imputation range for lhoggilt: Lo = 135.0000 Up = 537.5000
lhoggilt imputed using lhoggilt/ lhgexp46 ratio

Bounds computed using reported and imputed values

Fields	Revised	Reported	Lower	Upper
lhgfdpur	0.000	0.000	0.000	18000.000
lhpgsld1	1032.000	1032.000	590.909	1545.454
lhpgsld2	927.000	927.000	638.182	1669.091
lhpgsld3	911.000	911.000	531.250	1625.000
lhoggilt	233.000	978.000	135.000	537.500
lhg180br	7123.000	7123.000	2162.060	100000000.000
lhg60brd	3716.000	3716.000	911.000	100000000.000
lhgexp13	125.000	125.000	54.276	233.386
lhgexp46	135.000	135.000	48.848	233.386
lhgpgsld	2870.000	2870.000	1155.000	5197.500
lhgfdcst	0.000	0.000	0.000	200.278
lhgfdlbs	0.000	0.000	0.000	120.000
lhgfar13	385.000	385.000	212.593	956.667
lhgpig13	2870.000	2870.000	0.007	7123.000
lhogpig3	911.000	911.000	0.004	3716.000
lhogtotl	7123.000	7123.000	2870.000	100000000.000
lhogfar1	125.000	125.000	83.471	218.308
lhogfar2	135.000	135.000	74.978	196.096
lhogfar3	125.000	125.000	70.077	214.353
dumone	1.000	1.000	0.000	100000000.000
lhgund60	2710.000	2710.000	0.000	100000000.000
lhgto119	1490.000	1490.000	0.000	100000000.000
lhgto179	1413.000	668.000	0.000	100000000.000

lhgov180	1249.000	1249.000	0.000	100000000.000
lhogboar	28.000	28.000	0.000	100000000.000
lhogpig1	1032.000	1032.000	0.000	100000000.000
lhogpig2	927.000	927.000	0.000	100000000.000
lhogsld1	0.000	0.000	0.000	100000000.000
lhogsld2	0.000	0.000	0.000	100000000.000
lhogsld3	0.000	0.000	0.000	100000000.000

Record 52 had two failed ratio edits and at least one induced edit failed. The solution to the error localization problem identified the values of the variables lhgto179 and lhoggilt to be deleted and imputed. The value for the variable lhoggilt was imputed using ratio regression. The value imputed for the variable lhgto179 was used to satisfy a balance edit (This is known because no imputation range was calculated for the variable lhgto179). The action taken by the SPEER edit system for this record can be summarized as reallocating 745 hogs from sows and gilts to the market hog category of up to 179 pounds.

The output file sum.out provides useful information about the frequency of failed edits and the frequency of the variable values deleted. Listed below is the output of the file sum.out.

Central values read for 1 category
1155 records were run thru the SPEER edit

Ratio lhgfdpur/ lhgfdcst	failed 1	times
Ratio lhpgsld1/ lhogfar1	failed 15	times
Ratio lhpgsld2/ lhogfar2	failed 14	times
Ratio lhpgsld3/ lhogfar3	failed 8	times
Ratio lhoggilt/ lhgexp13	failed 11	times
Ratio lhoggilt/ lhgexp46	failed 9	times
Ratio lhg180br/ lhgpig13	failed 1	times
Ratio lhg60brd/ lhogpig3	failed 7	times
Ratio lhgexp13/ lhgexp46	failed 4	times
Ratio lhgpgsld/ lhgfar13	failed 2	times
Ratio lhgfdcst/ lhgfdlbs	failed 16	times
Ratio lhgfdlbs/ dumone	failed 2	times
Ratio lhgpig13/ lhogtotl	failed 5	times

Field lhpgsld1	was deleted 5 times
Field lhpgsld2	was deleted 9 times
Field lhpgsld3	was deleted 4 times
Field lhoggilt	was deleted 16 times
Field lhg60brd	was deleted 7 times

Field lhgexp13 was deleted 1 times
Field lhgexp46 was deleted 3 times
Field lhgfdcst was deleted 1 times
Field lhgfdlbs was deleted 18 times
Field lhgfar13 was deleted 5 times
Field lhgpig13 was deleted 5 times
Field lhogpig3 was deleted 2 times
Field lhogtotl was deleted 10 times
Field lhogfar1 was deleted 14 times
Field lhogfar2 was deleted 11 times
Field lhogfar3 was deleted 9 times
Field lhgund60 was deleted 2 times
Field lhgto119 was deleted 1 times
Field lhgto179 was deleted 4 times
Field lhgov180 was deleted 3 times
Field lhogboar was deleted 6 times
Field lhogpig1 was deleted 1 times
Field lhogpig2 was deleted 3 times
Field lhogsld1 was deleted 1 times
Field lhogsld2 was deleted 3 times

The above output can be somewhat misleading and should be interpreted only as a guide due to modifications made to the SPEER edit system. Since data records were run iteratively through the main (spr3d.for) program, it is possible that a ratio edit failed more than once or that a variable value was deleted more than once for a given record. Thus, the above output can overstate the number of times a ratio edit failed and the number of times a variable value was deleted. This output does, however, provide feedback about the ratio edits and the actions taken by the SPEER edit system.