



**AgEcon** SEARCH  
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search  
<http://ageconsearch.umn.edu>  
[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

**EVALUATION OF THE AGGIES AUTOMATED 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, January 1999, NASS Research Report, Report No. RD-99-01.

## **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 or a batch 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 consistent manner.

This paper evaluates a generalized automated edit and imputation system developed by the author called the Agricultural Generalized Imputation and Edit System (AGGIES). The AGGIES is appealing for the following reasons: 1) editing and imputation are fully automated, 2) the system provides consistency in the edit and imputation process, and 3) the system can be easily applied to any number of surveys, thus conserving resources to the development and maintenance of a single system. Comparisons between the AGGIES and the current edit and imputation procedures are made for expanded totals and the number and magnitude of variable changes. The data used for these comparisons are obtained from the Quarterly Hog Survey. The results reveal that the expanded totals obtained from using the AGGIES are similar to those obtained from the current edit and imputation procedures. Further testing on more applications is recommended.

## **KEY WORDS**

Generalized Automated Edit and Imputation System; Error Localization; Imputation Estimators

The views expressed herein are not necessarily those of NASS or USDA. This report was prepared for limited distribution to the research community outside the U.S. Department of Agriculture.

## **ACKNOWLEDGMENTS**

The author would like to thank the Iowa SSO for providing Key-Entry III data files for use in this study; Roberta Pense for helpful guidance throughout the project and for reviewing drafts of this report; and Kara Perritt for reviewing drafts of this report. Thanks to management (Ron Bosecker and George Hanuschak) for their support of this project. Special thanks to Robert Hood for help with SAS/AF and to Kara Broadbent for help with SAS/IML.

## TABLE OF CONTENTS

<b>SUMMARY</b> .....	iii
<b>1. INTRODUCTION</b> .....	1
<b>2. OVERVIEW OF THE AGGIES</b> .....	3
<b>3. AGGREGATE LEVEL STATISTICS</b> .....	7
<b>4. CONCLUSIONS AND RECOMMENDATIONS</b> .....	11
<b>REFERENCES</b> .....	13
<b>APPENDIX 1—RESULTS FROM FIVE RUNS</b> .....	15
<b>APPENDIX 2—DETAILS ON THE SYSTEM</b> .....	26
<b>INITIATING SYSTEM</b> .....	26
<b>EDIT SPECIFICATION</b> .....	28
<b>SPECIFYING EDITS IN THE AGGIES</b> .....	29
<b>FORMATION OF EDIT/DATA GROUPS</b> .....	31
<b>CHECK EDITS</b> .....	34
<b>EDIT SUMMARY</b> .....	35
<b>OUTLIER DETECTION</b> .....	35
<b>ERROR LOCALIZATION</b> .....	37
<b>GENERAL DESCRIPTION</b> .....	37
<b>MATHEMATICAL FORMULATION</b> .....	39
<b>IMPUTATION</b> .....	41
<b>EVALUATION OF THE AGGIES</b> .....	46
<b>EDIT SPECIFICATION</b> .....	46
<b>FORMATION OF EDIT/DATA GROUPS</b> .....	49
<b>APPENDIX 3—VARIABLE NAMES</b> .....	51

## 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 Census of Agriculture. After data collection and prior to the summarization and publication of statistics, the data are edited for completeness and consistency. Obtaining data that are accurate is important for making inference of the underlying population characteristics (e.g., estimating population totals and ratios). The 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 edit and imputation 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 edit and imputation process can often be time consuming, requiring considerable staff hours to complete the editing and imputation tasks. Hence, the data editing and imputation 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 and imputation 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, imputing and publishing survey data, new procedures are constantly sought to improve the edit and imputation process.

The Agricultural Generalized Imputation and Edit System (AGGIES) offers the potential to improve the efficiency of the edit and imputation process while also performing editing and imputation in a timely manner. The AGGIES is an automated edit and imputation system, developed by the author. 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 statistical frequency structure of the data file should be maintained.
- 3) The imputation rules should derive 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.

It is noted that NASS has systems which perform automated edit and automated imputation to some extent. These include the Crops & Stocks Survey imputation routine, zeroing out data during a survey edit, and the Census edits using the complex edit. However, the tasks of editing and

imputation have not been tied together as in a generalized automated edit and imputation system such as the AGGIES whereby the imputation rules are derived from the edit rules.

The AGGIES is comprised of a number of modules, each performing a separate function.

Edits are specified interactively in the edit specification module. The edits are required to be of linear form, linear inequalities and/or linear equalities. The edits are conditions that describe an acceptable record. Sometimes it may be desirable to apply different edits to different data records. For example, different edits may be applicable to the data records in different strata. This is accomplished by forming edit groups containing one or more edits and data groups containing one or more data records. For each edit group formed, a data group to which the edit group will be applied is formed.

Once the edits and/or edit groups have been specified, they should be checked for logical consistency, redundancy, determinacy and hidden equalities using the check edits module. Since the edits were required to be of linear form, this is easily accomplished using linear programming techniques (Giles, 1988). It is desired to have a minimal set of edits, resulting from the elimination of redundant edits and the identification of hidden equality edits, to avoid slowing the system processing in subsequent modules. The process of specifying edits, forming groups and checking the edits may become a cyclical process, being repeated several times until a final set of edits and/or edit groups is decided upon.

Upon the receipt of data, which are assumed to be continuous and non-negative, the results of applying the edits to the data can be observed with the selection of the edit summary module. This module displays for each edit specified along with positivity edits, the number of records passing and failing the edit. This summary can provide useful information about the edit set such as edits that are too restrictive or not restrictive enough. The outlier detection module compares a variable's value for a particular record with the value for all records in the file being edited for detecting outlying values. The use of this module provides an inter-record edit in addition to the intra-record edits specified in the edit set.

For data records failing one or more edits, the error localization module identifies, for each data record, the fewest values to change so that the data record can satisfy all of the edits. Weights can be assigned to the variables, in which case the module identifies, for each data record, the fewest weighted values to change so that the data record can satisfy all of the edits. Once the error localization module has been run, the values identified to be changed must be imputed so that each data record satisfies all edits. Prior to the actual imputation of values, the following information needs to be specified: 1) the order in which the variable values are to be imputed, 2) whether or not imputed values should contribute to the averages in the imputation estimators, and 3) which imputation estimators, if any, are to be applied to each variable and their order of application, if more than one is selected. Note that each data record is guaranteed to satisfy all specified edits after imputation.

The results from evaluating a subset of the September 1996 Iowa Quarterly Hog Report reveal that the expanded totals obtained from using the AGGIES are mostly similar to those obtained from the

current edit and imputation procedures, with the AGGIES making approximately sixty percent fewer changes. Of the twenty-one survey variables, thirteen (including all major survey indications) had average absolute expanded differences of less than one percent, five had average absolute expanded differences between one and five percent, two had average absolute expanded differences between five and ten percent, and one had an average absolute expanded difference exceeding ten percent. These results were obtained by the AGGIES in less than thirty minutes on a 233 Mhz Pentium computer.

Several recommendations are presented for the further evaluation of the AGGIES. In particular, it is recommended that the AGGIES be evaluated using Crops & Stocks Survey data, Census data and Sheep Survey data. Additionally, it is recommended that the imputation options be expanded to include donor imputation.

## 1. INTRODUCTION

The National Agricultural Statistics Service (NASS) conducts a wide variety of agricultural surveys. Among these are the Quarterly Agricultural Surveys which are used to collect current agricultural production data. Data collection begins around the first of each quarter; editing and imputation 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, imputing and publishing survey data, new and innovative procedures are sought to improve the efficiency while maintaining the timeliness of the edit and imputation process.

This paper describes the generalized automated edit and imputation system AGGIES (AGricultural Generalized Imputation and Edit System) which was developed by the author using SAS/IML and SAS/AF. A generalized automated edit and imputation system is a generalized system that receives as input a set of edits which describes an acceptable record. The system applies simultaneously the set of edits to each data record. A data record that does not satisfy the set of edits, because of missing or erroneous values, has a subset of its values changed according to some criterion so that the modified data record adheres to the set of edits. The system performs completely the editing and imputation tasks. There are several advantages associated with the use of a generalized automated edit and imputation system.

First, a generalized automated system provides for more efficient editing and imputation of the data with the potential for cost and time savings. The high relative cost associated with editing is documented in the report, Data Editing in Federal Statistical Agencies (1990). With the use of the computer to perform editing and imputation in a single system rather than the traditional way of manual editing and imputing, the timeliness of the edit and imputation process can be improved with a reduction in resources.

Second, the system provides consistency in the edit and imputation process. That is, the results of data records run through the system will be similar regardless of when or where they were edited and imputed. The editing and imputation are performed objectively with the results being nearly repeatable. Only when there are multiple solutions identified in the error localization module can the results differ when using the system, on different occasions, with the same edit and imputation specifications. However, with the assignment of variable weights in the error localization module, the variability between running the system on different occasions can be significantly reduced or even eliminated.

Third, an audit trail, which is the tracking of changes made to the data records and the reasons the changes were made, can be easily established and stored. 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.

Finally, a generalized automated edit and imputation system can be easily applied to any

number of surveys, thus conserving resources to the development and maintenance of a single system. With the AGGIES, survey specific linear edits that describe an acceptable record are written for each survey, but the error correction and imputation schemes applied are then derived from the input specifications (i.e., edits, order of variable imputation, and imputation estimators selected) and the data, rather than being specified for each possible outcome. Generalized systems may not allow as much flexibility as a survey-specific edit and imputation system. However, a survey-specific system, as its name implies, must be re-written for each survey which can consume a significant amount of resources. It is expected that the compromise in flexibility would be a minor issue when compared to the amount of resources required to develop a survey-specific edit and imputation system. An example of a survey-specific edit and imputation system is the complex edit which has been used to edit and impute for the U.S. Agricultural Censuses. Although the complex edit performs editing and imputation automatically, the variable values to change and the imputation of the values are specifically coded into the system via if-then statements based on the outcome of edits applied to each data record.

The development of the AGGIES emanated from a previous project in which the SPEER (Structured Programs for Economic Editing and Referrals) automated edit and imputation system was evaluated. In the research report entitled "Evaluation of the SPEER Automatic Edit and Imputation System" (Todaro, 1997), the shortcomings of SPEER were stated. These shortcomings limited its use as an editing and imputation tool in the NASS edit process. The primary limitation was that only a restricted set of edits could be specified to

the system because the error localization algorithm utilized was only useful for error localizing records when the edits were ratio edits and simple equality edits.

Statistics Canada's Generalized Edit and Imputation System (GEIS) was also not recommended in (Todaro, 1997) for use because of the software in which it was implemented. It appeared, however, that a system such as the GEIS would be useful to the NASS edit and imputation process. The GEIS is more general than SPEER in that the edits specified can be general linear edits, not merely ratio edits and simple equality edits which are a subset of general linear edits. Several members of the NASS Research Review Committee that reviewed the above mentioned research report expressed interest in the development of a generalized automated edit and imputation system. As a result, Research Division staff decided to develop a generalized automated edit and imputation system possessing many of the same features as the GEIS. The main advantage of developing a system is that it could be tailored to NASS's editing and imputation needs using software supported by NASS.

Section 2 of this report provides an overview of the functionality and the implementation of these modules, while Appendix 2 provides additional detail on the system as well as the theoretical and mathematical background used. In Section 3 the results of using the AGGIES to edit and impute data from the September 1996 Iowa Quarterly Hog Report will be presented. Section 4 discusses the conclusions and recommendations for future actions.

## **2. OVERVIEW OF THE AGGIES**



The use of a generalized automated edit and imputation system such as the AGGIES is based upon several assumptions (Morabito and Shields, 1992). The first assumption is that follow-up is complete; that is, no more attempts will be made to obtain incomplete data or to recontact respondents in the case of inconsistent or erroneous data. This decision could be based on the anticipated little gain in doing so, or because resources have been exhausted. Another assumption is that only those records with a lesser impact on the aggregate statistics are run through a generalized automated edit and imputation system. Since a generalized automated edit and imputation system always changes data that do not conform to the edits (i.e., no warning edits), data may be changed in an undesirable way for records that have a significant impact on aggregate statistics. Two final assumptions are that the data must be continuous and non-negative, and the edits must be of linear form.

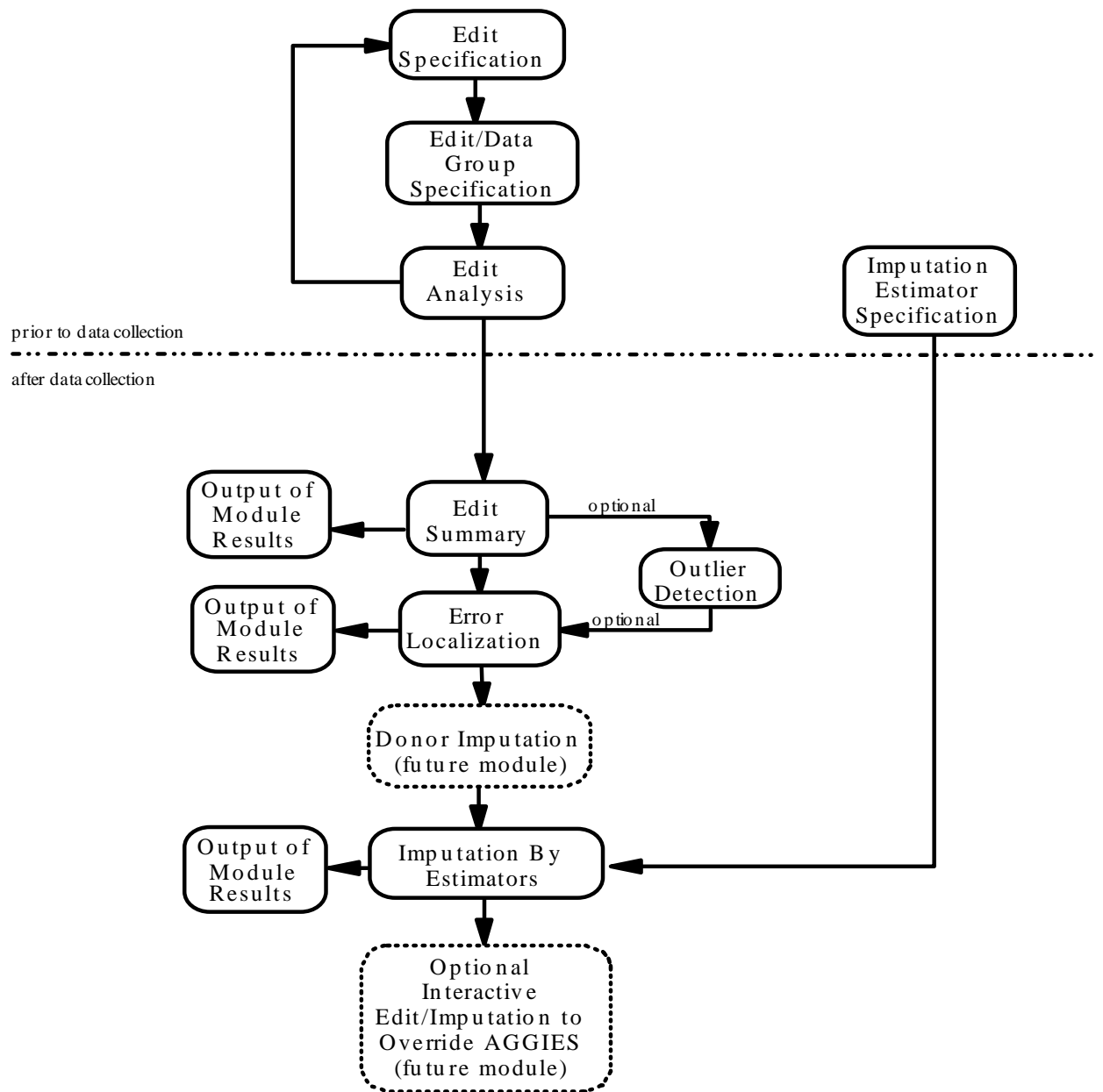
An interactive graphical user interface which allows for viewing and modifying changes made by the AGGIES needs to be incorporated. Upon review of the AGGIES edited and imputed data, the statistician may decide to make changes to some data records. The statistician could then opt to resubmit the data to the AGGIES or to override the system. In the case that the statistician decides to override the AGGIES, there needs to be a facility for keying comments. These comments would be useful for constructing an audit trail.

The AGGIES is comprised of a number of modules, each performing a separate function (See Figure 1). Initiating the system requires running the SAS set-up program 'aggies.sas'. Once this program is run, the file to be edited

and identification variables which uniquely identify the data records must be selected. The system then checks if the file selected has been edited previously using the AGGIES. Based on this check, different screens will be displayed to the user.

The first module allows the editor to specify the set of edits. It is the major input into the system. The edits are required to be of linear form, linear inequalities and/or linear equalities. Edits that are not linear, such as ratio and conditional edits, can often be recast as linear edits. If there are edits that cannot be written in linear form, then they must be applied outside of the system. The edits are the conditions that describe an acceptable record. If a record fails one or more edits because of missing or inconsistent data, the system will change the value of one or more variables in order to make the record satisfy all of the edits. Thus, the quality of the resulting imputed data set is directly affected by the edits.

Specifying edits as linear functions of the variables is somewhat different from the traditional manner of formulating edits (See Appendix 2; Evaluation of the AGGIES, Edit Specification). Traditionally edits have been formulated as if-then conditions. The if-condition acts as the edit while the then-condition specifies an action to take (imputation) or information about possible actions to take, in the form of an error message (i.e., editing and imputation are combined into a single statement).



**Figure 1.** AGGIES Flowchart

This method of editing and imputation can be very cumbersome and difficult to maintain because of the large number of if-then conditions required to describe an acceptable record. This approach to editing and imputation is also survey specific. Each survey requires a separate programming of a large number of if-then conditions.

Sometimes it may be desirable to apply different edits to different data records. For example, different edits may be applicable to the data records in different strata. This is accomplished by forming edit groups containing one or more edits and data groups containing one or more data records in the second module. For each edit group formed, a data group to which the edit group will be applied is formed.

An advantage of forming edit groups is that only those records satisfying the data group condition are used in imputing a variable value. This is true for imputation estimators using only data from the data set being edited and for imputation estimators using data from a historical data set.

Once the edits and/or edit groups have been specified, they should be checked for logical consistency, redundancy, determinacy and hidden equalities in the third module. Since the edits were required to be of linear form, this is easily done using linear programming theory (see Giles, 1988). These conditions are most easily described by using the region formed by the edit set called the feasible region. The edit set is inconsistent if the feasible region is empty. If the removal of an edit from the edit set leaves the feasible region unchanged, then that edit is a redundant edit. Determinacy occurs when the set of edits results in a variable whose value is required to equal a

single value. A hidden equality is an equality edit, not contained in the edit set, that is implied by one or more edits specified in the edit set.

Since the computationally intensive error localization algorithm uses the edits along with the data records to identify which values to change, it is desirable to identify a minimal set of edits representing the same feasible region described by the originally specified set of edits. The specification and checking of the edits should be performed in a cyclical manner. After the edits have been specified, they should be checked. If the result of checking the edits leads to the addition, deletion or modification of any edits, then the modified set of edits should again be checked. This will ensure that only the minimal set of edits will be processed by the system in subsequent modules.

The edit set can and should be specified and analyzed prior to the receipt of data. Once data have been collected, the edits can be applied to the data records. The fourth module, the edit summary module, displays for each edit contained in the edit set along with positivity edits, the number of records passing and failing the edit. This summary can provide useful information about the edit set (Cotton, 1993). First, the observation of edits with high rates of failure may reveal edits that were mis-specified and/or too restrictive. Second, edits with high failure rates may result in a high rate of imputation for certain variables, possibly affecting inferences made from the survey data. Third, since the error localization module can be time-consuming, the results can provide an estimate of the time this module will take to process the records. Finally, if there are variables whose values are required to be integers (e.g, livestock), this module could be run after rounding the imputed data values to

determine if the rounding caused any edit failures.

The outlier detection module compares a variable's value for a particular record with the value for all records in the file being edited for detecting outlying values. The use of this module provides an inter-record edit in addition to the intra-record edits specified in the edit set. This module has been programmed to identify those data records with an outlying value for a variable and in which the value is involved in at least one failed edit. This allows for the possibility of detecting large operations with inconsistencies among the reported values. It is undesirable for a generalized automated edit and imputation system to make large changes to such records as the aggregate statistics can be significantly altered. Rather, these records can be identified and reviewed manually.

The error localization module identifies, for each data record, the fewest values to change so that the record can satisfy all of the edits. This is intuitively appealing; changing the fewest values per record is equivalent to preserving as much of the originally reported data as possible. A weight, corresponding to the perceived reliability of the input data for a particular variable, can be assigned to each of the variables. In this case the module identifies, for each data record, the fewest weighted values to change so that the record can satisfy all of the edits (See Appendix 2; Error Localization). Several sets of values, all being minimal, may be identified. When this occurs, the module randomly selects a set. Occasionally, there are a few records that consume an inordinate amount of processing time in the error localization module. To avoid having a few records slow down the system, an option has been added which sets an upper

limit on the amount of processing time for a single record. If a record exceeds this time, it is identified for manual review and the data remain unchanged. The resulting output summaries from running this module consists of two parts. The first summary tabulates the number of times each variable value was identified to be deleted. The second summary lists, for data records identified to have at least one variable value deleted, the originally reported record followed by the same data record with values of minus one inserted for the values identified to be deleted.

Once the error localization module has been run, the deleted values must be imputed so that each data record satisfies all edits. Prior to the actual imputation of values, the following information needs to be specified : 1) the order in which the variable values are to be imputed, 2) whether or not imputed values should contribute to the averages in the imputation estimators, and 3) which imputation estimators, if any, are to be applied to each variable and their order of application for each variable. Up to six estimators may be specified for each variable. If none of the selected imputation estimators results in a value that will result in the record satisfying all edits, the set of values that results in the variable satisfying all edits is calculated, and the midpoint of this set is imputed. Thus after imputation, it is guaranteed that the record will satisfy all edits.

Imputation estimates are calculated based on imputation "batches". These imputation batches may be an edit batch or they may be multiple edit batches. At this point, there is no minimum number of observations required for calculating an imputation estimate. However, because the estimate must satisfy all edits, it should be reasonable.

The six available imputation estimators are (i denotes the unit, t the time period, x the auxiliary variable, and y the survey variable):

Current Mean - the mean of values in the file being edited.

$$y_{it} = \bar{y}_t$$

Current Ratio - an auxiliary variable adjusted by the ratio of the current mean to the mean of an auxiliary variable. Only those records that contribute to both averages are used in calculating the ratio.

$$y_{it} = \frac{\bar{y}_t}{\bar{x}_t} x_{it}$$

Previous Value - the value from a previous file for the same unit.

$$y_{it} = y_{i(t-1)}$$

Previous Mean - the mean of values from a previous file.

$$y_{it} = \bar{y}_{(t-1)}$$

Auxiliary Trend - the previous value for the unit being imputed adjusted by the ratio of a current auxiliary variable to the auxiliary variable from a previous file.

$$y_{it} = \frac{x_{it}}{x_{i(t-1)}} y_{i(t-1)}$$

Difference Trend - the previous value adjusted by the ratio of the current mean to the previous mean of the value being imputed. Only those records that contribute to both averages are used in calculating the ratio.

$$y_{it} = \frac{\bar{y}_t}{\bar{y}_{(t-1)}} y_{i(t-1)}$$

Any or all of the imputation estimators may be selected for the variables requiring imputation

(See Appendix 2; Imputation). If the first imputation estimator results in a value that will result in the record satisfying all edits, then that value is imputed. Otherwise, the next imputation estimator specified is considered. If none of the selected imputation estimators results in a value that will result in the record satisfying all edits, the set of values that results in the variable satisfying all edits is calculated, and the midpoint of this set is imputed.

The resulting output summaries from running the imputation module also consist of two parts. The first summary tabulates, for each imputation estimator, the number of times each variable was imputed. The second summary lists, for data records identified to have at least one variable value imputed, the originally reported record followed by the imputed data record. This second summary along with the second summary output from the running the error localization module provide useful information for constructing an audit trail.

A more thorough and complete explanation of the AGGIES, the theory behind it, the mathematical formulations, and examples of the functions can be found in Appendix 2.

### 3. AGGREGATE LEVEL STATISTICS

Aggregate statistics from the AGGIES are compared with those from the current Blaise/IDAS editing system, which is being treated as “truth” using the September 1996 Iowa Quarterly Hog Report. 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 and imputed 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 1.

Notice that the percentage of samples collected on paper is less than 27 percent in each of the three lowest strata. By contrast, at least 94 percent of the samples in each remaining stratum were collected on paper.

There were 1155 (subtract-level) records in the Key-Entry III file. Fifty-seven of these records were excluded from summary by the statistician during the survey because they were either not considered usable or because of adjustments made to compensate for frame

duplication.

The actual edits and edit groups for this analysis are listed in the last section of Appendix 2. They follow the recommendations of the Hog Edit and Analysis Team to the extent possible. For each variable, the midpoint of all values that result in satisfying all edits was used as the imputation option (i.e., no imputation estimators were specified).

In Section 2 it was mentioned that the error localization module randomly selects a solution set (a minimal set of variable values) when several sets, all being minimal, are identified. To assess the variability of the results obtained from the AGGIES, it was run five times. Table 3 in Appendix 1 contains the expanded totals from the current edit and imputation procedures and for each of the five runs, the difference of the expanded totals

**Table 1.** Population and Sample Size Counts by Stratum

Stratum	Population $N_h$	Sample selected $n_{hu}$	Sample collected on paper	% of sample collected on paper
80	4398	91	5	5.5
82	9283	366	53	14.5
84	7707	549	148	27.0
86	2922	419	397	94.7
88	950	314	300	95.5
92	161	121	117	96.7
98	25	25	25	100.0
Total	25446	1885	1045	

obtained from the AGGIES and the expanded totals from the current edit and imputation

procedures expressed as a percentage of the expanded totals from the current edit and

imputation procedures.

No variable weights were assigned to the variables. Viewing the results with no weights may provide some insight to assigning variable weights for subsequent runs.

There were no records in the five runs that exceeded the upper limit (30 minutes) on the amount of processing time for a single record in the error localization module. Normally, the upper limit would be no more than a few minutes. However, a large value for the upper limit was used to avoid, if possible, the occurrence of records requiring more processing time than specified by the upper limit. The time consumed for each of the five runs ranged from 13 minutes to 25 minutes on a 233 Mhz Pentium computer.

Three records were identified as outliers with respect to the total hog inventory (lhogtotl) variable in the outlier detection module. This variable was selected since its value provides

a reliable measure of size of an operation. These records would require editing and imputation by some other means. Since the final edited and imputed values were available, the Key-Entry III values were replaced with the final values (i.e., edited and imputed using the current procedures) for these three records. However, for some of these records, the AGGIES imputed the same values as the current procedures.

Table 2 displays the average expanded totals from running the data set five times through the AGGIES. Of the twenty-one variables (the bold entries in Table 2 are aggregate variables and are excluded), thirteen had average absolute expanded differences of less than one percent, five had average absolute expanded differences between one and five percent, two had average absolute expanded differences between five and ten percent, and one had an average absolute expanded difference exceeding ten percent.

**Table 2.** Comparison of Average Expanded Totals for Paper Collected Data Only

Variable	AGGIES Average	Current Procedures	Percentage Difference
Total Hogs & Pigs	7,305,365	7,306,858	-0.02
Market Hogs & Pigs under 60 LBS	2,039,314	2,048,913	-0.47
Market Hogs & Pigs 60-119 LBS	1,721,633	1,720,461	0.07
Market Hogs & Pigs 120-179 LBS	1,474,730	1,473,543	0.08
Market Hogs & Pigs 180+ LBS	1,357,564	1,355,212	0.17
Boars & Young Males for Breeding	31,114	28,960	7.44
Sows & Gilts for Breeding	681,010	679,770	0.18
Sows expected to farrow in next 3 mo.	351,625	351,357	0.08
Sows expected to farrow in 4-6 mo.	325,824	325,680	0.04
<b>Sows Farrowed the last 3 mo.</b>	<b>343,625</b>	<b>342,397</b>	<b>0.36</b>
Sows Farrowed 3 mo. Ago	117,140	117,733	-0.50
Sows Farrowed 2 mo. Ago	105,197	105,067	0.12

**Table 2.** Comparison of Average Expanded Totals for Paper Collected Data Only

Variable	AGGIES Average	Current Procedures	Percentage Difference
Sows Farrowed 1 mo. Ago	121,288	119,597	1.41
<b>Total pig crop from last 3 mo.</b>	<b>2,941,278</b>	<b>2,932,297</b>	<b>0.31</b>
<b>Pig crop on hand from last 3 mo.</b>	<b>2,640,309</b>	<b>2,628,169</b>	<b>0.46</b>
Pig crop on hand from 3 mo. ago	846,817	835,833	1.31
Pig crop on hand from 2 mo. ago	820,475	813,816	0.82
Pig Crop from last mo.	973,017	978,520	-0.56
<b>Pigs sold or slaughtered from last 3 mo.</b>	<b>300,969</b>	<b>304,128</b>	<b>-1.04</b>
Pigs sold or slaughtered from crop 3 mo. ago	148,478	161,940	-8.31
Pigs sold or slaughtered from crop 2 mo. ago	87,671	91,533	-4.22
Pigs sold or slaughtered from last mo. crop	64,820	50,655	27.96
Feeder Pig Lb.	19,922	19,681	1.22
Feeder Pig Price	17,818	17,759	0.33
Feeder Pigs Purchased	192,191	187,340	2.59

The relatively large percentage difference for boars, 7.44 percent, was attributed to the AGGIES changing the boar variable value for a single record in three of the five runs and changing the value for two records in one run. There were no changes made to the boar inventories in the current system. Examining the expanded totals for the boar inventory for the four runs in which the AGGIES made a change reveals that the expanded total ranged from 29,369 to 33,892 resulting in a percentage difference ranging from 1.41 percent to 17.03 percent (See Appendix 1, Table 3). Since the boar inventory is relatively small compared to other inventory variable values, a slight change can result in a moderate to large percentage difference between the two systems. The absolute percentage differences for the two systems ranged from 0.10 to 54.52 for the three pig crop sold or slaughtered variables (See Appendix 1, Table 3). However, when official estimates are set, the pigs sold or slaughtered are aggregated to a three month

total. For the three month aggregate of the pigs sold or slaughtered, the absolute percentage differences for the two systems ranged from 0.46 to 7.32. The average absolute percentage difference of the five runs was 1.04.

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

Tables 4 through 8 (See Appendix 1) show the frequency of records, for each of the five runs, that 1) were not changed in either the AGGIES or in the current edit and imputation system, 2) not changed in the AGGIES but changed in the current edit and imputation system, 3) changed in the AGGIES but not changed in the current edit and imputation system, and 4) changed in the AGGIES and



changed in the current edit and imputation system. The entries reveal that the two systems usually did not make changes to the same record variable values except for the average price per head and average weight per head variables. The current system made over sixty percent more changes than the AGGIES (165 to 103; these numbers exclude the three records that were identified as outliers. These three records account for 7 changes made by the current system. For each table among Tables 4 through 8, these numbers may be obtained by subtracting 7 from numbers obtained).

Tables 9 through 13 (See Appendix 1) show, for each of the five runs, 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 and/or imputation 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.

From the entries in Tables 9 through 13, 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 some editing bias or problems associated with the questionnaire.

#### **4. CONCLUSIONS AND RECOMMENDATIONS**

Using the AGGIES has several potential advantages for NASS surveys:

1) Commodity data editing and imputation are performed by the system resulting in an edited and imputed data set similar to that currently produced by NASS, as demonstrated using the 1996 Iowa Hog Survey data. This minimizes the need for manually reviewing and correcting the data records which, in turn, allows for more efficient ways of editing and imputing data with the potential for cost and time savings.

2) The system provides an audit trail (See Appendix 2 for a description of the 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 for improving future surveys.

3) The system allows for consistency in the edit and imputation process. The editing and imputation are performed objectively with the results being nearly repeatable. Only when there are multiple solutions identified in the error localization module can the results differ when using the system, on different occasions, with the same edit and imputation specifications. The difference in the expanded totals between runs was generally small as seen in Table 3.

4) The system can be easily applied to any number of surveys, thus conserving resources to the development and maintenance of a single system. The major input into the system are the edits, not which values to change and impute for each situation.

5) The system, developed by NASS and coded in SAS, can be tailored to NASS's editing and imputation needs using software supported by NASS. Additional features and modifications

can be easily incorporated. NASS's experience in SAS is quite extensive.

However, there are several issues to address when using the AGGIES for NASS surveys and the Agricultural Census:

1) The AGGIES will not perform all editing functions. The system is designed to edit and impute for continuous data. Thus, the editing of completion codes and data adjustment factors must be performed outside of the system. Additionally the edits specified to the system are required to be of linear form rather than the usual conditional edits (See Appendix 2; Specifying Edits in the AGGIES).

2) A plan as to how the system could be implemented in NASS's Agricultural Survey processing to form a complete edit strategy and system integration is needed. In particular, which editing and analysis tools (Blaise, IDAS, SPS, etc.) need to be applied and their order of application needs to be determined. Processing platforms also need to be addressed.

3) It is assumed that only those records with a lesser impact on the aggregate statistics are run through the system since the system always changes data that do not satisfy all edits. Thus, there needs to be a policy decision, and then a mechanism for identifying which records are to be processed in the AGGIES for surveys and censuses. An interactive graphical-user interface also needs to be developed to allow the statistician to view and make changes to the AGGIES edited and imputed data.

4) Using the AGGIES to edit and impute for one survey period and one state's hog survey data has been evaluated. However, since other surveys and the Census of Agriculture collect

other types of data, and perform different types of edits, the AGGIES needs to be evaluated using these data. In addition, States from different regions need to be evaluated.

Therefore, the following recommendations are made:

1) Evaluate the use of AGGIES on other non-livestock survey data. We have the December 1996 Iowa Crop & Stocks Survey data, which can be evaluated.

2) Evaluate sections of the Agricultural Census starting with the hog section data, by reformulating the current edits into linear edits, with the assistance of commodity experts. Evaluate the impact of the AGGIES on the Census data by comparing the AGGIES output to the data after Final Data Review.

3) As recommended in the report of the Sheep Editing and Analysis Team, work with this team to conduct a post-survey test of the AGGIES for the 1999 January Sheep Survey, using data from the four largest sheep producing states. This evaluation will 1) allow operational employees to be involved in formulating linear edits and 2) provide feedback on the implementation issues of the AGGIES in NASS's survey processing.

4) Expand imputation options to include donor imputation. Donor imputation, "borrowing" data values from another similar record, may better preserve the multi-variate structure of the data set and can be used, although to a limited extent, to impute for categorical variables.

## REFERENCES

- Anderson, C. et al. (1996), "Report of the Hog Editing and Analysis Team," unpublished documentation, National Agricultural Statistics Service, USDA, Washington D.C.
- Anderson, C. et al. (1998), "Report of the Sheep Editing and Analysis Team," unpublished documentation, National Agricultural Statistics Service, USDA, Washington D.C.
- Bazaraa, M.S., Jarvis, J.J., and Sherali, H.D. (1990), *Linear Programming and Network Flows*. Second Edition. NY: Wiley.
- Chernikova, N.V. (1964), "Algorithm for Finding a General Formula for the Nonnegative Solutions of a System of Linear Equations," *U.S.S.R. Computational Mathematics and Mathematical Physics*, No. 4, 151-158.
- Chernikova, N.V. (1965), "Algorithm for Finding a General Formula for the Nonnegative Solution of a System of Linear Inequalities," *U.S.S.R. Computational Mathematics and Mathematical Physics*, No. 5, 228-233.
- Cotton, C. (1993), "Functional Description of the Generalized Edit and Imputation System," Statistics Canada Technical Report.
- Data Editing in Federal Statistical Agencies (1990), Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology. Statistical Policy Working Paper 18.
- Fellegi, I.P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, No. 71, 17-35.
- Giles, P. (1988), "A Model for Generalized Edit and Imputation of Survey Data," *The Canadian Journal of Statistics*, No. 16, 57-73.
- Giles, P., and Patrick, C. (1986), "Imputation Options in a Generalized Edit and Imputation System," *Survey Methodology*, Vol. 12, No. 1, 49-60.
- Groves, R.M. (1989), *Survey Errors and Survey Costs*. NY: Wiley.
- Hidioglou, M.A., and Berthelot, J.M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys," *Survey Methodology*, No. 12, 73-83.
- Hood, R., and Apodaca, M. (1996), "Improving the Quality of Survey Data Through an Interactive Data Analysis System," *Proceedings of the Twenty-First Annual SAS Users Group International Conference*, 753-760.
- 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.
- Morabito, J., and Shields, M. (1992), "Generalized Edit and Imputation System Applications User's Guide," Statistics Canada Technical Report.
- Rubin, D.S. (1975), "Vertex Generation in Cardinality Constrained Linear Programs," *Operations Research*, No. 23, 555-565.

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

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 Section on Survey Research Methods*. American Statistical Association, Vol. 1, 166-171.

Todaro, T.A., (1997), "Evaluation of the SPEER Automatic Edit and Imputation System," National Agricultural Statistics Service, USDA, Washington D.C., RD Research Report No. RD-97-04.

## APPENDIX 1—RESULTS FROM FIVE RUNS

**Table 3.** Comparison of Expanded Totals for Paper Collected Data Only

Variable	Current Procedures	% Diff Run 1	% Diff Run 2	% Diff Run 3	% Diff Run 4	% Diff Run 5
Total Hogs & Pigs	7,306,858	0.02	0.01	-0.06	-0.02	-0.04
Market Hogs & Pigs under 60 LBS	2,048,913	-0.25	-0.53	-0.55	-0.55	-0.46
Market Hogs & Pigs 60-119 LBS	1,720,461	0.00	0.04	0.29	0.00	0.00
Market Hogs & Pigs 120-179 LBS	1,473,543	0.07	0.34	-0.17	0.34	-0.17
Market Hogs & Pigs 180+ LBS	1,355,212	0.26	0.19	0.10	0.16	0.16
Boars & Young Males for Breeding	28,960	1.41	13.50	5.23	0.00	17.03
Sows & Gilts for Breeding	679,770	0.20	-0.09	0.19	0.35	0.27
Sows expected to farrow in next 3 mo.	351,357	0.08	0.08	0.08	0.08	0.08
Sows expected to farrow in 4-6 mo.	325,680	0.04	0.04	0.04	0.04	0.04
<b>Sows Farrowed the last 3 mo.</b>	<b>342,397</b>	<b>0.55</b>	<b>0.54</b>	<b>0.22</b>	<b>0.91</b>	<b>-0.43</b>
Sows Farrowed 3 mo. Ago	117,733	-0.88	-0.18	0.09	-0.89	-0.66
Sows Farrowed 2 mo. Ago	105,067	0.70	-0.28	-0.11	0.55	-0.24
Sows Farrowed 1 mo. Ago	119,597	1.83	1.97	0.65	2.99	-0.37
<b>Total Pig crop from last 3 mo.</b>	<b>2,932,297</b>	<b>0.43</b>	<b>0.74</b>	<b>0.00</b>	<b>0.71</b>	<b>-0.35</b>
<b>Pig crop on hand from last 3 mo.</b>	<b>2,628,169</b>	<b>0.43</b>	<b>0.51</b>	<b>0.43</b>	<b>0.48</b>	<b>0.46</b>
Pig crop on hand from 3 mo. Ago	835,833	0.71	1.88	1.67	1.04	1.28
Pig crop on hand from 2 mo. Ago	813,816	1.22	0.66	0.20	1.00	1.00
Pig crop from last mo.	978,520	-0.47	-0.78	-0.44	-0.42	-0.70
<b>Pigs sold or slaughtered from last 3 mo.</b>	<b>304,128</b>	<b>0.46</b>	<b>2.68</b>	<b>-3.72</b>	<b>2.69</b>	<b>-7.32</b>
Pigs sold or slaughtered from crop 3 mo. Ago	161,940	-7.18	-7.34	-7.74	-10.44	-8.85
Pigs sold or slaughtered from crop 2 mo. Ago	91,533	-4.67	-6.25	1.27	-2.75	-8.70
Pigs sold or slaughtered from last mo. crop	50,655	34.22	50.87	0.10	54.52	0.10
Feeder Pig Lb.	19,681	1.22	1.22	1.22	1.22	1.22
Feeder Pig Price	17,759	0.33	0.33	0.33	0.33	0.33
Feeder Pigs Purchased	187,340	2.59	2.59	2.59	2.59	2.59

**Table 4.** Comparison of Same Record Changes - Run 1

Variable	AGGIES: No change Current: No change		AGGIES: No change Current: Change	AGGIES: Change Current: No change	AGGIES: Change Current: Change
	Value=0	Value>0			
Total Hogs & Pigs	289	791	11	5	2
Market Hogs & Pigs under 60 LBS	445	639	10	1	3
Market Hogs & Pigs 60-119 LBS	431	656	9	0	2
Market Hogs & Pigs 120-179 LBS	440	648	8	1	1
Market Hogs & Pigs 180+ LBS	420	662	12	3	1
Boars & Young Males for Breeding	531	566	0	1	0
Sows & Gilts for Breeding	509	580	7	1	1
Sows expected to farrow in next 3 mo.	532	560	6	0	0
Sows expected to farrow in 4-6 mo.	571	520	6	0	1
Sows Farrowed 3 mo. Ago	584	507	0	7	0
Sows Farrowed 2 mo. Ago	627	469	0	2	0
Sows Farrowed 1 mo. Ago	590	506	1	1	0
Pig crop on hand from 3 mo. Ago	627	459	9	2	1
Pig crop on hand from 2 mo. Ago	647	442	6	3	0
Pig crop from last mo.	595	490	4	7	2
Pigs sold or slaughtered from crop 3 mo. Ago	1019	67	8	4	0
Pigs sold or slaughtered from crop 2 mo. Ago	1053	38	4	3	0
Pigs sold or slaughtered from last mo. crop	1067	24	4	3	0
Feeder Pig Lb./Head	1001	78	1	1	17
Feeder Pig \$/Head	1001	79	1	1	16
Feeder Pigs Purchased	1000	79	2	1	16
Sum of all variables	13979	8860	109	47	63

**Table 5.** Comparison of Same Record Changes - Run 2

Variable	AGGIES: No change Current: No change		AGGIES: No change Current: Change	AGGIES: Change Current: No change	AGGIES: Change Current: Change
	Value=0	Value>0			
Total Hogs & Pigs	289	794	10	2	3
Market Hogs & Pigs under 60 LBS	445	639	11	1	2
Market Hogs & Pigs 60-119 LBS	431	654	9	2	2
Market Hogs & Pigs 120-179 LBS	439	649	8	1	1
Market Hogs & Pigs 180+ LBS	420	665	10	0	3
Boars & Young Males for Breeding	530	566	0	2	0
Sows & Gilts for Breeding	509	581	7	0	1
Sows expected to farrow in next 3 mo.	532	560	6	0	0
Sows expected to farrow in 4-6 mo.	571	520	6	0	1
Sows Farrowed 3 mo. Ago	584	512	0	2	0
Sows Farrowed 2 mo. Ago	627	469	0	2	0
Sows Farrowed 1 mo. Ago	590	506	1	1	0
Pig crop on hand from 3 mo. Ago	627	457	8	4	2
Pig crop on hand from 2 mo. Ago	647	441	5	4	1
Pig crop from last mo.	595	491	5	6	1
Pigs sold or slaughtered from crop 3 mo. Ago	1018	66	6	6	2
Pigs sold or slaughtered from crop 2 mo. Ago	1054	37	4	3	0
Pigs sold or slaughtered from last mo. crop	1066	25	4	3	0
Feeder Pig Lb./Head	1001	78	1	1	17
Feeder Pig \$/Head	1001	79	1	1	16
Feeder Pigs Purchased	1000	79	2	1	16
Sum of all variables	13976	8868	104	42	68

**Table 6.** Comparison of Same Record Changes - Run 3

Variable	AGGIES: No change Current: No change		AGGIES: No Change Current: Change	AGGIES: Change Current: No change	AGGIES: Change Current: Change
	Value=0	Value>0			
Total Hogs & Pigs	289	794	11	2	2
Market Hogs & Pigs under 60 LBS	445	640	11	0	2
Market Hogs & Pigs 60-119 LBS	430	656	9	1	2
Market Hogs & Pigs 120-179 LBS	440	648	8	1	1
Market Hogs & Pigs 180+ LBS	420	665	10	0	3
Boars & Young Males for Breeding	531	566	0	1	0
Sows & Gilts for Breeding	509	579	7	2	1
Sows expected to farrow in next 3 mo.	532	560	6	0	0
Sows expected to farrow in 4-6 mo.	571	520	6	0	1
Sows Farrowed 3 mo. Ago	584	512	0	2	0
Sows Farrowed 2 mo. Ago	627	470	0	1	0
Sows Farrowed 1 mo. Ago	590	504	0	3	1
Pig crop on hand from 3 mo. Ago	626	456	5	6	5
Pig crop on hand from 2 mo. Ago	647	443	5	2	1
Pig crop from last mo.	595	491	4	6	2
Pigs sold or slaughtered from crop 3 mo. Ago	1019	68	7	3	1
Pigs sold or slaughtered from crop 2 mo. Ago	1051	37	4	6	0
Pigs sold or slaughtered from last mo. crop	1069	25	4	0	0
Feeder Pig Lb./Head	1001	78	1	1	17
Feeder Pig \$/Head	1001	79	1	1	16
Feeder Pigs Purchased	1000	79	2	1	16
Sum of all variables	13977	8870	101	39	71



**Table 7.** Comparison of Same Record Changes - Run 4

Variable	AGGIES: No change Current: No change		AGGIES: No change Current: Change	AGGIES: Change Current: No change	AGGIES: Change Current: Change
	Value=0	Value>0			
Total Hogs & Pigs	289	791	10	5	3
Market Hogs & Pigs under 60 LBS	445	640	11	0	2
Market Hogs & Pigs 60-119 LBS	431	655	9	1	2
Market Hogs & Pigs 120-179 LBS	439	649	8	1	1
Market Hogs & Pigs 180+ LBS	420	665	12	0	1
Boars & Young Males for Breeding	532	566	0	0	0
Sows & Gilts for Breeding	508	579	7	3	1
Sows expected to farrow in next 3 mo.	532	560	6	0	0
Sows expected to farrow in 4-6 mo.	571	520	6	0	1
Sows Farrowed 3 mo. Ago	584	507	0	7	0
Sows Farrowed 2 mo. Ago	627	468	0	3	0
Sows Farrowed 1 mo. Ago	590	506	1	1	0
Pig crop on hand from 3 mo. Ago	627	456	8	5	2
Pig crop on hand from 2 mo. Ago	646	443	4	3	2
Pig crop from last mo.	595	493	5	4	1
Pigs sold or slaughtered from crop 3 mo. Ago	1020	68	8	2	0
Pigs sold or slaughtered from crop 2 mo. Ago	1053	38	4	3	0
Pigs sold or slaughtered from last mo. crop	1065	25	4	4	0
Feeder Pig Lb./Head	1001	78	1	1	17
Feeder Pig \$/Head	1001	79	1	1	16
Feeder Pigs Purchased	1000	79	2	1	16
Sum of all variables	13976	8865	107	45	65

**Table 8.** Comparison of Same Record Changes - Run 5

Variable	AGGIES: No change Current: No change		AGGIES: No change Current: Change	AGGIES: Change Current: No change	AGGIES: Change Current: Change
	Value=0	Value>0			
Total Hogs & Pigs	289	792	11	4	2
Market Hogs & Pigs under 60 LBS	445	640	10	0	3
Market Hogs & Pigs 60-119 LBS	431	656	9	0	2
Market Hogs & Pigs 120-179 LBS	440	648	8	1	1
Market Hogs & Pigs 180+ LBS	420	665	10	0	3
Boars & Young Males for Breeding	532	565	0	1	0
Sows & Gilts for Breeding	508	578	7	4	1
Sows expected to farrow in next 3 mo.	532	560	6	0	0
Sows expected to farrow in 4-6 mo.	571	520	6	0	1
Sows Farrowed 3 mo. Ago	584	510	0	4	0
Sows Farrowed 2 mo. Ago	627	468	0	3	0
Sows Farrowed 1 mo. Ago	590	504	0	3	1
Pig crop on hand from 3 mo. Ago	627	455	8	6	2
Pig crop on hand from 2 mo. Ago	647	441	6	4	0
Pig crop from last mo.	595	492	5	5	1
Pigs sold or slaughtered from crop 3 mo. Ago	1020	68	7	2	1
Pigs sold or slaughtered from crop 2 mo. Ago	1054	38	3	2	1
Pigs sold or slaughtered from last mo. crop	1069	25	4	0	0
Feeder Pig Lb./Head	1001	78	1	1	17
Feeder Pig \$/Head	1001	79	1	1	16
Feeder Pigs Purchased	1000	79	2	1	16
Sum of all variables	13983	8861	104	42	68

**Table 9.** Comparison of the Direction of Changes - Run 1

Variable	AGGIES			Current System		
	Positive	Negative	No Change	Positive	Negative	No Change
Total Hogs & Pigs	6	1	1091	9	4	1085
Market Hogs & Pigs under 60 LBS	4	0	1094	11	2	1085
Market Hogs & Pigs 60-119 LBS	2	0	1096	6	5	1087
Market Hogs & Pigs 120-179 LBS	2	0	1096	5	4	1089
Market Hogs & Pigs 180+ LBS	1	3	1094	5	8	1085
Boars & Young Males for Breeding	1	0	1097	0	0	1098
Sows & Gilts for Breeding	2	0	1096	6	2	1090
Sows expected to farrow in next 3 mo.	0	0	1098	2	4	1092
Sows expected to farrow in 4-6 mo.	0	1	1097	2	5	1091
Sows Farrowed 3 mo. Ago	1	6	1091	0	0	1098
Sows Farrowed 2 mo. Ago	1	1	1096	0	0	1098
Sows Farrowed 1 mo. Ago	0	1	1097	0	1	1097
Pig crop on hand from 3 mo. ago	1	2	1095	3	7	1088
Pig crop on hand from 2 mo. ago	2	1	1095	2	4	1092
Pig crop from last mo.	1	8	1089	2	4	1092
Pigs sold or slaughtered from crop 3 mo. ago	4	0	1094	7	1	1090
Pigs sold or slaughtered from crop 2 mo. ago	3	0	1095	3	1	1094
Pigs sold or slaughtered from last mo. crop	2	1	1095	3	1	1094
Feeder Pig Lb./Head	0	18	1080	0	18	1080
Feeder Pig \$/Head	1	16	1081	0	17	1081
Feeder Pigs Purchased	0	17	1081	0	18	1080

**Table 10.** Comparison of the Direction of Changes - Run 2

Variable	AGGIES			Current System		
	Positive	Negative	No Change	Positive	Negative	No Change
Total Hogs & Pigs	5	0	1093	9	4	1085
Market Hogs & Pigs under 60 LBS	3	0	1095	11	2	1085
Market Hogs & Pigs 60-119 LBS	3	1	1094	6	5	1087
Market Hogs & Pigs 120-179 LBS	2	0	1096	5	4	1088
Market Hogs & Pigs 180+ LBS	1	2	1095	5	8	1085
Boars & Young Males for Breeding	2	0	1096	0	0	1098
Sows & Gilts for Breeding	1	0	1097	6	2	1090
Sows expected to farrow in next 3 mo.	0	0	1098	2	4	1092
Sows expected to farrow in 4-6 mo.	0	1	1097	2	5	1091
Sows Farrowed 3 mo. Ago	0	2	1096	0	0	1098
Sows Farrowed 2 mo. Ago	0	2	1096	0	0	1098
Sows Farrowed 1 mo. Ago	0	1	1097	0	1	1097
Pig crop on hand from 3 mo. ago	3	3	1092	3	7	1088
Pig crop on hand from 2 mo. ago	2	3	1093	2	4	1092
Pig crop from last mo.	0	7	1091	2	4	1092
Pigs sold or slaughtered from crop 3 mo. ago	6	2	1090	7	1	1090
Pigs sold or slaughtered from crop 2 mo. ago	3	0	1095	3	1	1094
Pigs sold or slaughtered from last mo. crop	3	0	1095	3	1	1094
Feeder Pig Lb./Head	0	18	1080	0	18	1080
Feeder Pig \$/Head	1	16	1081	0	17	1081
Feeder Pigs Purchased	0	17	1081	0	18	1080

**Table 11.** Comparison of the Direction of Changes - Run 3

Variable	AGGIES			Current System		
	Positive	Negative	No Change	Positive	Negative	No Change
Total Hogs & Pigs	3	1	1094	9	4	1085
Market Hogs & Pigs under 60 LBS	2	0	1096	11	2	1085
Market Hogs & Pigs 60-119 LBS	3	0	1095	6	5	1087
Market Hogs & Pigs 120-179 LBS	1	1	1096	5	4	1089
Market Hogs & Pigs 180+ LBS	1	2	1095	5	8	1085
Boars & Young Males for Breeding	1	0	1097	0	0	1098
Sows & Gilts for Breeding	2	1	1095	6	2	1090
Sows expected to farrow in next 3 mo.	0	0	1098	2	4	1092
Sows expected to farrow in 4-6 mo.	0	1	1097	2	5	1091
Sows Farrowed 3 mo. Ago	1	1	1096	0	0	1098
Sows Farrowed 2 mo. Ago	0	1	1097	0	0	1098
Sows Farrowed 1 mo. Ago	1	3	1094	0	1	1097
Pig crop on hand from 3 mo. ago	7	4	1087	3	7	1088
Pig crop on hand from 2 mo. ago	0	3	1095	2	4	1092
Pig crop from last mo.	1	7	1090	2	4	1092
Pigs sold or slaughtered from crop 3 mo. ago	4	0	1094	7	1	1090
Pigs sold or slaughtered from crop 2 mo. ago	6	0	1092	3	1	1094
Pigs sold or slaughtered from last mo. crop	0	0	1098	3	1	1094
Feeder Pig Lb./Head	0	18	1080	0	18	1080
Feeder Pig \$/Head	1	16	1081	0	17	1081
Feeder Pigs Purchased	0	17	1081	0	18	1080

**Table 12.** Comparison of the Direction of Changes - Run 4

Variable	AGGIES			Current System		
	Positive	Negative	No Change	Positive	Negative	No Change
Total Hogs & Pigs	6	2	1090	9	4	1085
Market Hogs & Pigs under 60 LBS	2	0	1096	11	2	1085
Market Hogs & Pigs 60-119 LBS	2	1	1095	6	5	1087
Market Hogs & Pigs 120-179 LBS	2	0	1096	5	4	1089
Market Hogs & Pigs 180+ LBS	0	1	1097	5	8	1085
Boars & Young Males for Breeding	0	0	1098	0	0	1098
Sows & Gilts for Breeding	4	0	1094	6	2	1090
Sows expected to farrow in next 3 mo.	0	0	1098	2	4	1092
Sows expected to farrow in 4-6 mo.	0	1	1097	2	5	1091
Sows Farrowed 3 mo. Ago	1	6	1091	0	0	1098
Sows Farrowed 2 mo. Ago	1	2	1095	0	0	1098
Sows Farrowed 1 mo. Ago	1	0	1097	0	1	1097
Pig crop on hand from 3 mo. ago	3	4	1091	3	7	1088
Pig crop on hand from 2 mo. ago	2	3	1093	2	4	1092
Pig Crop from last mo.	0	5	1093	2	4	1092
Pigs sold or slaughtered from crop 3 mo. ago	2	0	1096	7	1	1090
Pigs sold or slaughtered from crop 2 mo. ago	3	0	1095	3	1	1094
Pigs sold or slaughtered from last mo. crop	4	0	1094	3	1	1094
Feeder Pig Lb./Head	0	18	1080	0	18	1080
Feeder Pig \$/Head	1	16	1081	0	17	1081
Feeder Pigs Purchased	0	17	1081	0	18	1080

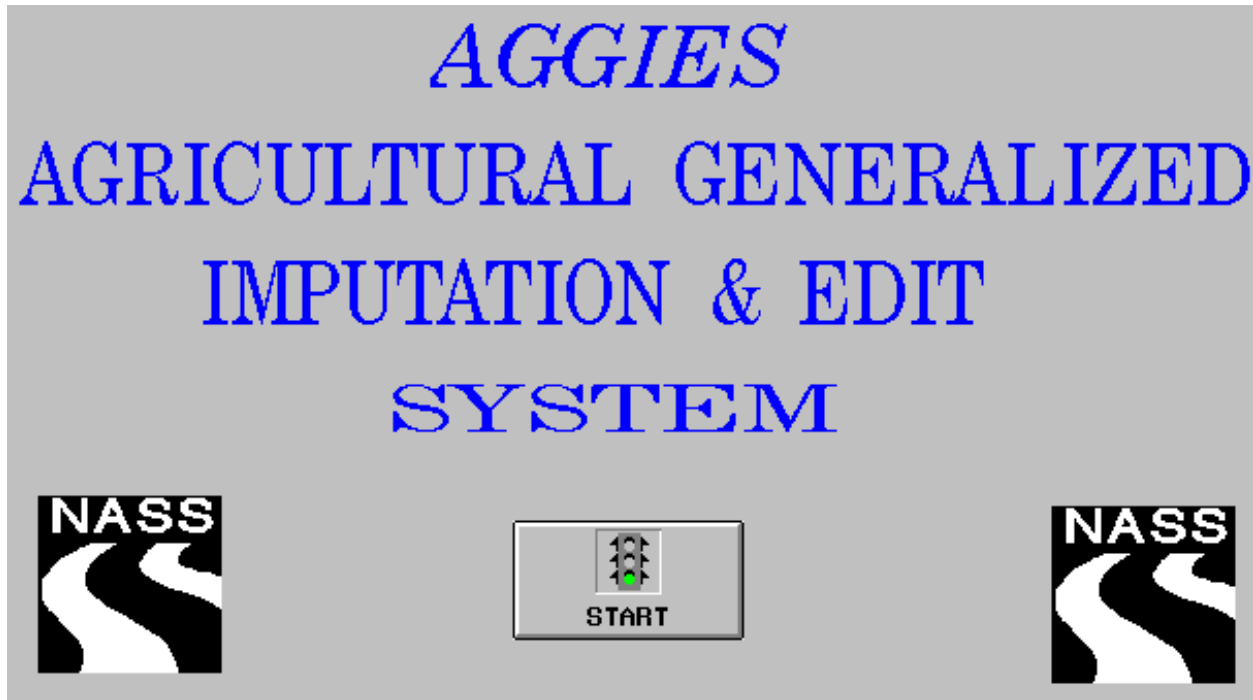
**Table 13.** Comparison of the Direction of Changes - Run 5

Variable	AGGIES			Current System		
	Positive	Negative	No Change	Positive	Negative	No Change
Total Hogs & Pigs	5	1	1092	9	4	1085
Market Hogs & Pigs under 60 LBS	3	0	1095	11	2	1085
Market Hogs & Pigs 60-119 LBS	2	0	1096	6	5	1087
Market Hogs & Pigs 120-179 LBS	1	1	1096	5	4	1089
Market Hogs & Pigs 180+ LBS	1	2	1095	5	8	1085
Boars & Young Males for Breeding	1	0	1097	0	0	1098
Sows & Gilts for Breeding	4	1	1093	6	2	1090
Sows expected to farrow in next 3 mo.	0	0	1098	2	4	1092
Sows expected to farrow in 4-6 mo.	0	1	1097	2	5	1091
Sows Farrowed 3 mo. Ago	0	4	1094	0	0	1098
Sows Farrowed 2 mo. Ago	0	3	1095	0	0	1098
Sows Farrowed 1 mo. Ago	0	4	1094	0	1	1097
Pig crop on hand from 3 mo. ago	4	4	1090	3	7	1088
Pig crop on hand from 2 mo. ago	1	3	1094	2	4	1092
Pig Crop from last mo.	0	6	1092	2	4	1092
Pigs sold or slaughtered from crop 3 mo. ago	3	0	1095	7	1	1090
Pigs sold or slaughtered from crop 2 mo. ago	2	1	1095	3	1	1094
Pigs sold or slaughtered from last mo. crop	0	0	1098	3	1	1094
Feeder Pig Lb./Head	0	18	1080	0	18	1080
Feeder Pig \$/Head	1	16	1081	0	17	1081
Feeder Pigs Purchased	0	17	1081	0	18	1080

## APPENDIX 2–DETAILS ON THE SYSTEM

### INITIATING SYSTEM

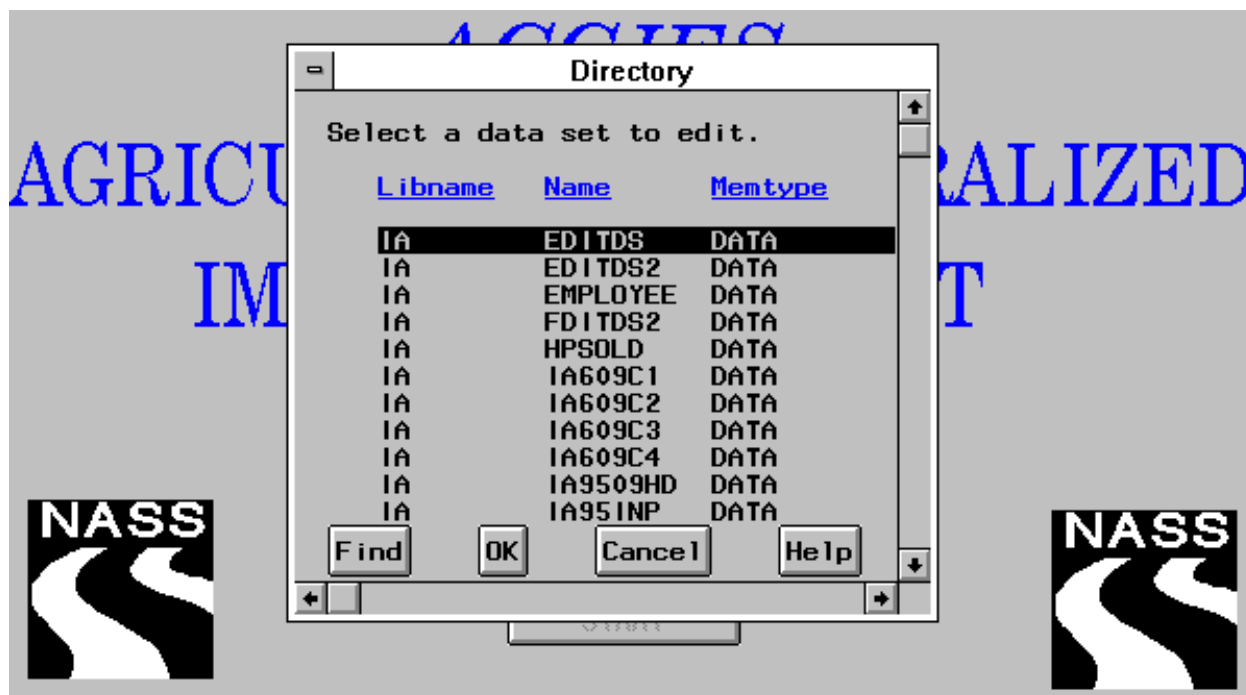
The AGGIES is initiated by running the set-up program ‘aggies.sas’. This results in Figure 1 being displayed.



**Figure 1.** Initial Screen of the AGGIES

Clicking on the ‘Start’ icon displays a listing of SAS data files as shown in Figure 2. The SAS data files, listed for this illustration, are located in the SAS library ‘IA’ associated with the directory ‘f:\users\todato\’. The SAS data files to be displayed can come from any number of directories. These directories can be included in the set-up program by associating them with SAS libraries.





**Figure 2.** Listing of SAS Data Files

This screen is used for selecting a data file to edit. Note that the “OK” push button does not need to be clicked. Once a file is selected, the next screen, shown in Figure 3, appears immediately.

A listing of SAS variable names is displayed. Up to five identification variables whose values uniquely identify each data record may be selected. This information will be needed in the imputation module if certain imputation estimators are selected, namely, previous value, auxiliary trend, or difference trend. It is also used to identify the data records in the summary outputs. Clicking on a variable in the list selects it as an identification variable. After a variable has been selected, an asterisk is placed to the left of the variable. Once selected, a variable can be de-selected by clicking on it in the list, thus removing the asterisk. The set of identification variables are submitted by clicking on the “OK” push button.



**Figure 3.** Selection of Identification Variables

After the identification variables have been selected, the system checks to see if this data file has an edit specification file associated with it (i.e., edit specifications already exist). If it has an edit specification file, the next screen to appear is the utility screen shown in Figure 5, otherwise the next screen to appear is the edit specification screen shown in Figure 4.

### EDIT SPECIFICATION

This module entails specifying the conditions which describe an acceptable record. Edit specification is the major input into a generalized automated edit and imputation system. The edits form the foundation of such a system which affect the imputation process. As a result the imputed data set can only be as good as the edits specified. Edits can be formulated by analyzing the questionnaire, performing data analysis or by using subject matter expertise (Morabito and Shields, 1992).

The conditions or edits are required to be of linear form :  $A_1X_1=b_1$  and/or  $A_2X_2\#b_2$ , where  $A_1$  is an  $m_1 \times n_1$  matrix of coefficients,  $A_2$  is an  $m_2 \times n_2$  matrix of coefficients,  $A=[A_1^T|A_2^T]^T$  is an  $m \times n$  ( $m=m_1+m_2$ ,  $n=n_1+n_2$ ) matrix of coefficients,  $X=[X_1^T|X_2^T]^T$  is an  $n \times 1$  vector of variable values,  $b=[b_1^T|b_2^T]^T$  is an  $m \times 1$  vector of constants,  $m$  is the number of edits and  $n$  is the number of variables involved in the edits. This requirement is imposed because many of the algorithms used to process the data are based on linear programming theory. A record that does not satisfy an edit is said to fail the edit; one that does is said to pass the edit.

Specifying edits as linear functions of the variables is somewhat different from the traditional manner of formulating edits (See Appendix 2; Evaluation of the AGGIES, Edit Specification). Traditionally edits have been formulated as if-then conditions. The if-condition acts as the edit while the then-condition specifies an action to take (imputation) or information about possible actions to take, in the form of an error message (i.e., editing and imputation are combined into a single statement).

This method of editing and imputation can be very cumbersome and difficult to maintain because of the large number of if-then conditions required to describe an acceptable record. This approach to editing and imputation is also survey specific. Each survey requires a separate programming of a large number of if-then conditions.

### SPECIFYING EDITS IN THE AGGIES

Linear edits are specified by entering an edit identifier, the coefficients of the variables, the variables, a relational operator and a constant. The edit specification screen is shown in Figure 4.

Enter Edit Identifier :

COEFFICIENT			COEFFICIENT		
<input type="text"/>	X1	_____	<input type="text"/>	X6	_____
<input type="text"/>	X2	_____	<input type="text"/>	X7	_____
<input type="text"/>	X3	_____	<input type="text"/>	X8	_____
<input type="text"/>	X4	_____	<input type="text"/>	X9	_____
<input type="text"/>	X5	_____	<input type="text"/>	X10	_____

**Figure 4.** Edit Specification Screen

The edit identifiers can be up to 8 characters in length and will be used as aliases for the associated edits in later modules of the system. The coefficients of the variables are typed in the rectangular region to the left of the push buttons X1-X10 and must be numeric. If an invalid value is entered, the system prompts the user to enter a valid numeric value. Clicking on one of the ten push buttons displays the list of names of the numeric variables in the data set being edited. A variable name is selected by clicking on the variable name in the list, after which, it appears to the right of the

associated push button. Up to ten variable names may be used to construct an edit (If needed, this can be increased). The variable names that form the edits must be entered sequentially beginning with push button X1. The system will not accept the selection of a variable name associated with a push button unless variables have been selected for all lower numbered push buttons. Each variable may only be used once per linear edit. For example,  $0.5 * \text{lhogtotl} + 0.5 * \text{lhogtotl} > 1$  would not be allowed since the variable `lhogtotl` was used twice. A relational operator is selected by clicking on the push button “Rel”. A choice can be selected from “<=”, “=”, and “>=”. Note the operators “<” and “>” are not available since it is assumed that the data being edited are continuous. The constant (or the right hand side of the edit) is entered by typing a numeric value in the rectangular region to the right of the text “Constant”. If the value entered is not numeric, the system prompts the user to enter a valid numeric value. The edit specification screen can be cleared by clicking on the “Undo Edit” push button. Once an edit has been entered, it can be submitted to the system by clicking on the “Submit Edit” push button. This saves the edit and clears the screen at which time another edit may be entered. Finally, the system may be exited by selecting the “Quit” push button.

Once the edits have been specified, the system allows for the viewing of all edits, modifying edits, deleting edits, adding more edits and other options by clicking on the “Continue” push button which displays the following screen, Figure 5.



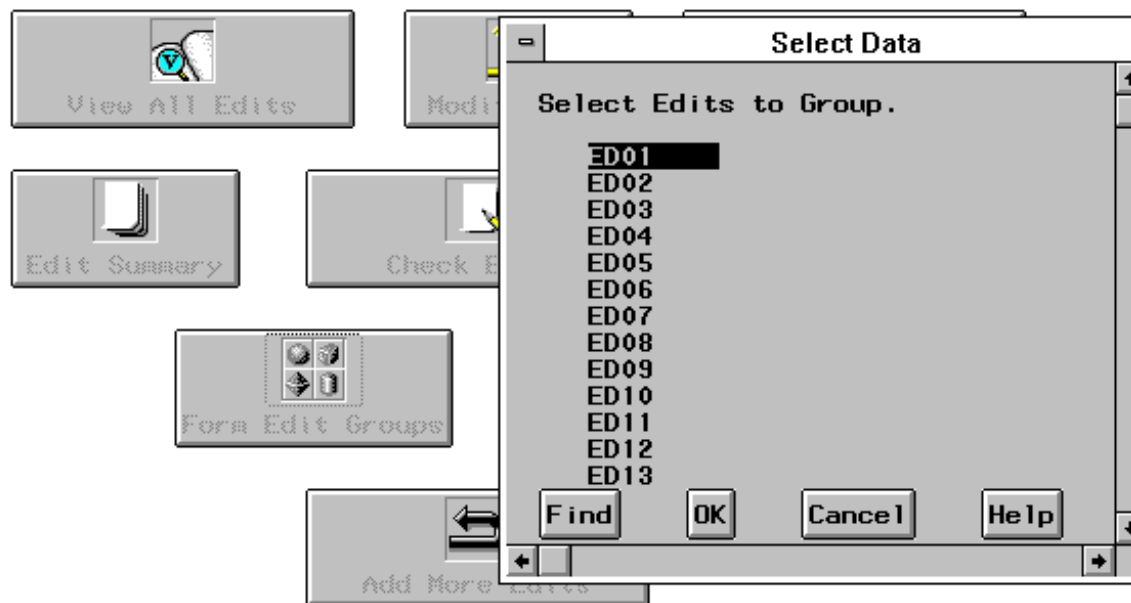
**Figure 5.** Utility Screen

Selecting the “View All Edits” icon displays each edit identifier and associated edit. The edits may be modified by selecting the “Modify Edit” icon. Selecting this option displays the list of edit identifiers for edits that have been entered into the system for the data set being edited. The selection of a particular edit identifier displays the edit specification screen (Figure 4) with the values filled in for the edit selected to be modified. An edit may be deleted by selecting the “Delete Edit” icon which

displays the list of edit identifiers for edits that have been entered into the system. A particular edit is deleted by selecting the corresponding edit identifier. Additional edits may be specified by selecting the “Add More Edits” icon. The selection of this option displays the edit specification screen.

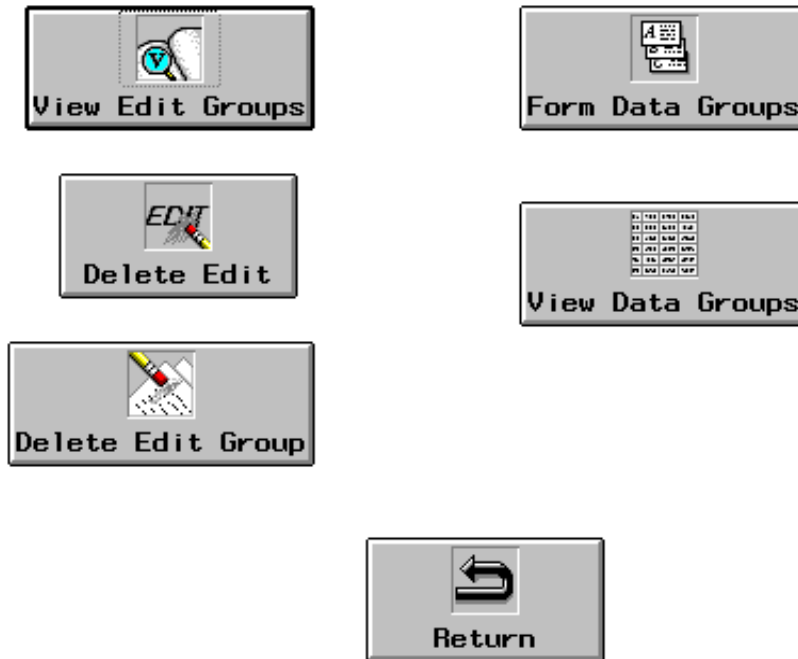
## FORMATION OF EDIT/DATA GROUPS

The AGGIES includes a module for applying a set of edits, called an edit group, to a certain group of data records, called a data group. For example, the data groups may be the data records belonging to the strata from sample design: one data group for each stratum. Because each stratum may have unique properties, a different set of edits may be required for each. For each stratum, an edit group is formed and its edits applied to the data records belonging to the stratum. An edit group can be specified by clicking on the “Form Edit Groups” icon in Figure 5. Clicking on this icon displays the list of edit identifiers for the edits that have been entered into the system as shown in Figure 6.



**Figure 6.** Display of Edit Identifiers

An edit group is formed by selecting all of the associated edit identifiers that will belong to the group. The selection of an edit identifier is made by clicking on it in the list. After an edit identifier has been selected, an asterisk is placed to the left of the edit identifier. Once selected, an edit identifier can be de-selected by clicking on it in the list, thus removing the asterisk. An edit group is submitted by clicking on the “OK” push button. (Note : Clicking on the “OK” push button without selecting any edit identifiers does not form an edit group. This can be done to view existing edit/data groups without creating any additional groups.) This results in the following screen, Figure 7, being displayed.



**Figure 7.** Edit and Data Group Screen

Selecting the icon “View Edit Groups” displays three columns : Edit Group, Edit Identifier, and Edit. The Edit Group column contains the edit group numbers assigned in ascending order beginning with 1 up to the number of edit groups formed. The edit identifier column contains the edit identifiers associated with the edits forming the edit group. Finally, the third column contains the actual edits forming the edit group.

The edit groups can be modified by clicking on the “Delete Edit” icon or the “Delete Edit Group” icon. The “Delete Edit” icon allows for the deletion of a particular edit in a particular edit group. Selecting this icon displays the list of edit groups that have been created. Selecting a particular edit group displays the list of edit identifiers associated with the edit group. Clicking on an edit identifier deletes the associated edit from the edit group. The “Delete Edit Group” icon allows for the deletion of an entire edit group. Clicking on this icon displays the list of edit groups that have been created. Selecting a particular edit group deletes all edits within the group and removes the edit group.

A data group is formed by specifying a condition which describes data records to be included in the group. For example, a data group can be formed by specifying the condition `stratum=85` or `lhogtotl>500`. Clicking on the “Form Data Groups” icon displays the following screen, Figure 8.

The screenshot shows a software interface for forming data groups. At the top, there is a button labeled "Edit Group" next to a small rectangular input field. Below this, the text "Where :" is followed by a long horizontal rectangular input field. Underneath the input field, there is a row of five buttons: "Variable", "Operator", "Sign", "Conjunction", and "Undo". Below these buttons, there are two larger buttons: "Submit" (which contains a stick figure icon) and "Return" (which contains a curved arrow icon).

**Figure 8.** Formation of Data Groups Screen

Clicking on the “Edit Group” push button lists all edit group numbers. The selection of an edit group number specifies that the edit group is to be applied to the data records belonging to the data group formed using this screen. To facilitate the specification of a data group, several push buttons have been provided. The “Variable” push button, when clicked, displays the list of the names of the numeric variables in the data set being edited. When a variable name is selected from this list, it appears in the rectangular region to the right of the text “Where :”. The next three push buttons described are operations which are used in forming the data group. Clicking the “operator” push button displays the list of operators “/”, “\*”, “-”, and “+”. The “Sign” push button, when clicked, displays the following list of operators “<”, “>”, “<=”, “=”, “>=”, and “^=”. The last operator, “^=”, has the meaning “not equal”. The “Conjunction” push button displays the list containing “AND” and “OR”, when clicked. The selection of an operator results in the operator being appended to the end of the equation in the rectangular region. A number used in forming the data group must be typed in the rectangular region. The equation can contain up to 200 characters. The final push button, “Undo”, clears the contents in the rectangular region. A data group is submitted by clicking on the “Submit” icon. Clicking on the “Return” icon returns to the previous screen shown in Figure 7.

The “View Data Groups” icon (Figure 7) displays two columns, Edit Group and Data Group. The Edit Group column contains the edit group numbers. The Data group column contains the equation that forms the data group (e.g.,  $lhogtotl > 500$ ). If no data group is specified, the entire data file is used as the data group.

An advantage of forming groups is that only those records satisfying the data group condition are used in imputing a variable value. This is true for imputation estimators using only data from the data

set being edited and for imputation estimators using data from a historical data set. For example, suppose the data records in each stratum were defined as a group and the current mean was selected as an imputation estimator. This situation allows using stratum means to impute within stratum rather than the mean of the entire data file.

## **CHECK EDITS**

This module is selected by clicking on the “Check Edits” icon in Figure 5. The edits specified are analyzed by edit/data group using linear programming theory (See for example, Bazaraa et al. (1990)). More specifically, the edits are checked for logical consistency, redundancy, hidden equalities, and determinacy. These conditions are most easily described by using the region formed by the edits called the acceptable (feasible) region. A record whose values lie within this region results in the record satisfying all edits. Otherwise, the record fails one or more edits. If the acceptable region is not empty, the set of edits is said to be logically consistent. If an edit, after being removed, results in the same acceptance region, then it is redundant. Hidden equalities are equality edits that are not explicitly specified, but rather implied by two or more edits that have been explicitly specified. Determinacy occurs when the set of edits results in a variable whose value is required to equal a single value. The occurrence of determinacy may be the result of the edits being too restrictive.

These conditions are checked by solving a series of linear programs (See Giles, 1988). Checking for logical consistency requires solving a single linear program. The objective function can be any linear function of the variables in the edits, say the first edit, with the constraints being the set of edits. If the acceptance region is non-empty, then the set of edits is logically consistent. A series of linear programs must be solved, in two steps, when checking for redundancy. In the first step, each edit is maximized subject to all of the edits. If any of the objective function values are non-zero, then the edit that is being maximized is redundant. The second step requires maximizing each edit subject to all edits except for the edit being maximized. If any of the objective function values are equal to zero, then the edit that is being maximized is redundant (The edit is a tight edit but redundant). Determinacy involves maximizing and minimizing each variable subject to the set of edits. If the maximum and minimum values are equal for a particular variable, then determinacy has occurred. Once all redundant edits have been removed, a check for hidden equalities can be performed. This check involves minimizing each edit subject to the set of non-redundant edits. If any of the objective function values are equal to zero, then the edit that is being minimized is an edit contributing to a hidden equality.

If logical inconsistency is detected, the output consists of a set of edits that, if removed, will result in a consistent set of edits. However, these edits should not be removed without a careful analysis of the entire set of edits. If redundant edits are detected, they are listed in the output of this module. If two or more edits can be rewritten as an equality edit, the edits that together imply the equality edit are displayed. Determinacy can be detected by viewing the variable ranges produced by this module. If the minimum and maximum values are equal for any variable, determinacy has occurred. When determinacy occurs for all variables, a message is displayed in the output noting that determinacy has occurred.

## **EDIT SUMMARY**



Clicking on the “Edit Summary” icon in Figure 5 selects this module. This module is the first to use data from the file being edited. Once the edits have been decided upon, they can now be applied to the data records. The output of this module displays, for each edit, the number of records that pass the edit and the number of records that fail the edit. The results are shown for both positivity edits and the user-specified edits (edit identifiers are displayed). The positivity edits are implied by the use of linear programming theory which requires the variable values to be non-negative. The failure rates of the positivity edits can be used to ascertain the amount of missing data since NASS uses “-1” to indicate a missing value. (Note, using the SAS missing value “.” will result in an error when using this module. Therefore, the value “.” should be replaced by “-1” or any other negative value to indicate a missing value.)

The results of this module provide an array of useful information for the statistician (Cotton, 1993). First, the observation of edits with high rates of failure may reveal edits that were mis-specified and/or too restrictive. Second, edits with high failure rates may result in a high rate of imputation for certain variables, possibly affecting inferences made from the survey data. Third, since the error localization module can be time-consuming, the results can provide an estimate of the time this module will take to process the records. Finally, if there are variables whose values are required to be integers (e.g, livestock), this module could be run after rounding the imputed data values to determine if the rounding caused any edit failures.

## OUTLIER DETECTION

This module compares a variable’s value for a particular record with the value for all records in the file being edited for detecting outlying values. Such a comparison is referred to as a statistical edit, in contrast to micro-editing. Micro-editing compares a variable’s value to other values within the same record according to the relationships specified in the edits. The addition of a statistical-edit module allows greater flexibility in the editing process. It is noted that the entire data set is used for determining outlying values as opposed to determining outlying values by edit/data group. The purpose of this module is to identify large values that may have an unusually large impact on aggregate statistics. It is undesirable for a generalized automated edit and imputation system to make large changes to such records as the aggregate statistics can be significantly altered. Rather, these records can be identified and reviewed manually.

The methodology used is based on a technique described by Hidioglou and Berthelot (1986). The method used in this module is described as follows.

First, the quantities  $d_{Q1}$  and  $d_{Q3}$  are calculated for each variable of interest.

$$d_{Q1} = \text{Max}(M - Q1, |A * M|)$$

$$d_{Q3} = \text{Max}(Q3 - M, |A * M|)$$

M is the median, Q1 is the first quartile, Q3 is the third quartile, and A is referred to as a minimum distance multiplier used to ensure a minimum value for  $d_{Q1}$  and  $d_{Q3}$ . The quantity  $d_{Q1}$  represents the distance from Q1 to M while the quantity  $d_{Q3}$  represents the distance from M to Q3.

Second, the following quantities are calculated as

$$\text{Lower Bound} = M - C d_{Q1}$$

$$\text{Upper Bound} = M + C d_{Q3}$$

where C is a constant multiplier. If the value of a variable lies below the Lower Bound or exceeds the Upper Bound, then that value is considered an outlier.

This outlier detection module differs from the same module used in the GEIS in the following ways. First, only the current method and not the historical method is used. The historical method compares the ratio of a variable's current value to its previous value for the same unit to bounds based on the ratios for all records (with a current and previous value). Second, the GEIS allows for the outlying values to be imputed whereas this system does not. In the GEIS, this module is run prior to the error localization module to ensure that a minimum number of values is changed per record. However, since the AGGIES does not allow for outlying values to be imputed, this module could be run before or after the error localization module. Third, this program only displays an outlying value if it is also involved in at least one failed edit. This allows for the possibility of detecting large operations with inconsistencies among the reported values.


The IDAS (Interactive Data Analysis System; Hood and Apodaca, 1996), also coded in SAS, can be used as a macro-edit tool to detect outliers from the AGGIES edited and imputed file. The purpose of the outlier analysis module is to determine those larger records that would require more detailed editing, not to replace the more comprehensive macro-edits used in the IDAS.


Clicking on the "Outlier Detection" icon in Figure 5 selects the outlier detection module and displays the following screen, Figure 9.


Variable:

Variables	
LHGEXP13	↑
LHGEXP46	
LHGFARM1	
LHGFARM2	
LHGFARM3	↓
←	→

Enter Coefficient

 View Outliers

 Submit

 Return

### **Figure 9. Outlier Detection Screen**

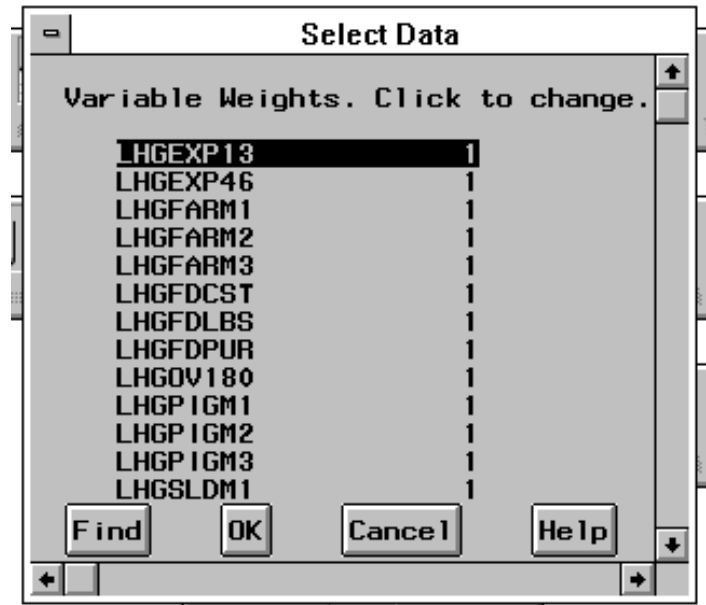
The constant multiplier,  $C$ , is entered in the rectangular region below the text “Enter Coefficient”. A value of six has been recommended as an initial value for  $C$  (Morabito and Shields, 1992). Based on the number of outlying records, the value can be modified. A default value of 0.05 is used for  $A$  (Morabito and Shields, 1992). A variable name is selected from the list of names of the numeric variables in the data set being edited by clicking on a variable name from the “Variables” listbox. Clicking on a variable name results in the variable name appearing in the rectangular region to the right of the text “Variable:”. The “View Outliers” icon, when clicked, displays those observations for which the selected variable value is a calculated outlier involved in one or more failed edits. Selecting the “Return” icon returns to the screen displayed in Figure 7. The “Submit” icon was added with the intent of imputing outlying values. This may be a future option in the AGGIES.

## **ERROR LOCALIZATION**

### **GENERAL DESCRIPTION**

The method of editing and imputation in a generalized edit and imputation system does not require the explicit specification of which values to change or the values to assign for a record that fails edits. The system controls what values to change and the values assigned based on some criterion. The criterion used for this system is to change the fewest values per record for a record failing edits. This criterion is intuitively appealing; changing the fewest values per record is equivalent to preserving as much of the originally reported respondent data as possible.

This module identifies, for each record, the minimal set of values to change in order for the record to satisfy all edits. Values are changed because either the record failed one or more edits or the record contained missing data. By assigning a negative value for the missing variables, the implied positivity edits would be violated. If a record fails at least one edit, some values must be changed for the record to adhere to the edits. Records with missing data would be changed to non-negative values. This module is selected by clicking on the “Error Localization” icon in Figure 5 which displays the following screen, Figure 10.



**Figure 10.** Specification of Variable Weights

Variable weights may be assigned for each variable involved in one or more user-specified edits. The default variable weights are one. If variable weights are assigned, the error localization module identifies, for each record, the minimal weighted set of values to change in order for the record to satisfy all edits. Thus, the higher the variable weight assigned, the less likely the variable value will be changed. The variable weights can be used to assign degrees of reliability to the variable values. A higher variable weight signifies more confidence in the values for that variable. Morabito and Shields (1992) discuss practical applications of using variable weights.

The solution obtained by the error localization module is not necessarily unique. Several sets of values, all being minimal, may be identified. When this occurs, the module randomly selects a set. A ramification of randomly selecting a minimal set is that the results may be different when running the module on different occasions, thus affecting the repeatability of the results. However, with the assignment of variable weights in the error localization module, the variability between running the system on different occasions can be significantly reduced or even eliminated.

There are two output summaries resulting from running this module. The first summary displays, for each edit/data group, the number of times each variable value was identified to be changed. After reviewing the results of this output, the module could be re-run with different variable weights for experimentation purposes. The second summary displays, for each record having one or more values identified to be changed, the original data record followed by the error localized data record. The distinguishing feature of the error localized record is the assignment of the value minus one to the values identified to be changed. This second output is useful for establishing an audit trail.

Occasionally, there are a few records that consume an inordinate amount of processing time in the error localization module. This occurs despite taking steps to make the underlying algorithm as efficient as possible. Statistics Canada has documented that a few records can consume the majority of processing time in the GEIS (Cotton, 1993). To avoid having the few records slow the system down, an option has been added into the code implementing the algorithm that sets an upper limit on the amount of processing time for a single record. Generally, if a record has not been processed within five minutes, it can consume a large amount of processing time. If the processing of a record exceeds this upper limit, the record observation number and the identification variable values for the record are printed in the output of the error localization module. For these records, the values for the original data record are the same as the error localized data record values.

## MATHEMATICAL FORMULATION

Given the set of edits  $\{x | Ax \# b, x \geq 0\}$  where  $A$  is an  $m \times n$  matrix of coefficients,  $x = (x_1, x_2, \dots, x_n)^T$  is the vector of variable values, and  $b = (b_1, b_2, \dots, b_m)^T$  is a vector of constants. The criterion in the error localization module is to change the minimum number of variable values per record so that after imputation, all edits will be satisfied. The approach taken is to add positive ( $y = (y_1, y_2, \dots, y_n)^T$ ) correction values and subtract negative ( $z = (z_1, z_2, \dots, z_n)^T$ ) correction values in making the change to these values. Both positive and negative corrections are needed since all elements of the correction vectors are assumed to be non-negative as required by linear programming theory. After the corrections have been applied, the variable values are represented as  $x + y - z$ . Thus, the set of edits can be rewritten as  $\{(y; z) | A(x + y - z) \# b, x + y - z \geq 0, y \geq 0, z \geq 0\}$ . Note that the variables are  $y$  and  $z$  and that  $x$  is known. The criterion can be restated as to minimize the number of non-zero corrections ( $y - z$ ). To be more formal, consider a cardinality function defined for each  $x = (x_1, x_2, \dots, x_n)$  by

$$f(x) = \sum_{i=1}^n w_i d(x_i) \quad \text{where} \quad d(x_i) = \begin{cases} 0 & \text{if } x_i = 0 \\ 1 & \text{otherwise} \end{cases} \quad i=1, \dots, n$$

where  $w_i$  is the positive weight coefficient associated with the  $i$ th variable. Thus, we would like to minimize  $f(y - z)$  subject to  $\{(y; z) | A(x + y - z) \# b, x + y - z \geq 0, y \geq 0, z \geq 0\}$ . Note that the vector  $(y - z)$  is in  $R^n$  while there are  $2n$  variables in the set of edits  $(y; z)$ . However, minimizing  $f(y - z)$  subject to the edits is equivalent to minimizing  $f(y; z)$  (Schiopu-Kratina and Kovar, 1989). Now, the problem can be restated in  $R^{2n}$ . This problem has been referred to as a cardinality constrained linear program. The solution to this problem can always be found in the set of vertices of the edit set (Rubin, 1975). An algorithm developed by Chernikova (Chernikova, 1964, 1965) is used to find the vertices. Actually, Chernikova's algorithm calculates all the edges of a convex polyhedral cone in the non-negative orthant with vertex at the origin. But, by the following lemma, the vertices of the region formed by the edit set can be found.

Lemma.  $\bar{x}$  is a vertex of  $F = \{x | Ax \# b, x \geq 0\}$  if and only if  $\{(\bar{x}^T, ? \geq 0)\}$  is an edge of the cone  $C_F = \{(x^T, ?)^T | Ax \# b, x \geq 0, ? \geq 0\}$ . Here  $?$  and  $?$  are scalar variables.

Note that interest is only in those solutions with  $\theta=1$ . The edit set can be rewritten into the corresponding equation of a cone as:

$$C_E = \{((y; z)^T, \theta)^T \mid -A(y-z) + (b - Ax)\theta \leq 0, y-z+x\theta \leq 0, y \geq 0, z \geq 0, \theta \leq 0\}.$$

Consider the associated matrix formed by  $C_E$ ,  $(D^T, I^{2n+1})^T$  where  $D = \begin{bmatrix} A & A & b - Ax \\ I^n & I^n & x \end{bmatrix}$  is an  $(m+n) \times (2n+1)$  matrix and  $I^{2n+1}$  is the  $(2n+1) \times (2n+1)$  identity matrix. Chernikova's algorithm, through a series of transformations, generates all edges of a cone transforming the matrix  $Y = (U^T, L^T)^T$  into a matrix  $\bar{Y}$  at each iteration, with the initial matrix being  $(D^T, I^{2n+1})^T$ . Although the matrices  $Y$  and  $\bar{Y}$  will have the same number of rows,  $\bar{Y}$  may have more or less columns. For  $w \in \mathbb{R}^{(2n+1)}$ , let  $w$  denote the ray  $\{\theta w, \theta \leq 0\}$ . The algorithm is given as follows:

- 0.1 If any row  $U$  has all components negative, then  $w=0$  is the only point in  $C_E$ .
- 0.2 If all the elements of  $U$  are non-negative, then the columns of  $L$  are the edges of  $C_E$ , i.e., the ray  $(l_j) = \{w = \theta l_j, \theta \leq 0\}$  is an edge of  $C_E$ ; here  $l_j$  denotes the  $j$ th column of  $L$ .
1. Choose the first row of  $U$ , say row  $r$ , with at least one negative element.
2. Let  $R = \{j \mid y_{rj} \leq 0\}$ . Let  $v = |R|$ , i.e., the number of elements of  $R$ . Then the first  $v$  columns of the new matrix,  $\bar{Y}$ , are all the  $y_j$  for  $j \in R$ , where  $y_j$  denotes the  $j$ th column of  $Y$ .
- 2'. If  $Y$  has only two columns and  $y_{r1}y_{r2} < 0$ , adjoin the column  $|y_{r2}|y_1 + |y_{r1}|y_2$  to the  $\bar{Y}$  matrix. Go to step 4.
3. Let  $S = \{(s, t) \mid y_{rs}y_{rt} < 0, s < t\}$ , i.e., the set of all (unordered) pairs of columns of  $Y$  whose elements in row  $r$  have opposite signs. Let  $I_0$  be the index of all non-negative rows of  $Y$ . For each  $(s, t) \in S$ , find all  $i \in I_0$  such that  $y_{is} = y_{it} = 0$ . Call this set  $I_1(s, t)$ . We now use some of the elements of  $S$  to create additional columns for  $\bar{Y}$ :
  - 3a. If  $I_1(s, t) = \emptyset$  (the empty set), then  $y_s$  and  $y_t$  do not contribute another column to the new matrix.
  - 3b. If  $I_1(s, t) \neq \emptyset$ , check to see if there is a  $u$  not equal to either  $s$  or  $t$ , such that  $y_{iu} = 0$  for all  $i \in I_1(s, t)$ . If such a  $u$  exists, then  $y_s$  and  $y_t$  do not contribute another column to the new matrix. If no such  $u$  exists, then choose  $a_1, a_2 > 0$  to satisfy  $a_1 y_{rs} + a_2 y_{rt} = 0$ . (One such choice is  $a_1 = |y_{rt}|, a_2 = |y_{rs}|$ .) Adjoin the column  $a_1 y_s + a_2 y_t$  to the new matrix.
4. When all pairs in  $S$  have been examined, and the additional columns (if any) have been added, we say that row  $r$  has been 'processed.' Now let  $Y$  denote  $\bar{Y}$  produced in processing row  $r$  and return to step 0.1.

When transforming the matrix  $Y$  to  $\bar{Y}$ , the number of columns can increase. To limit this increase, the rows are processed according to suggestions from Schiopu-Kratina and Kovar (1989). Rows corresponding to failed equality edits are processed first, failed inequality edits second, followed by equality edits (that pass) and then the remaining rows. Within each of these groups, the row to process is selected by calculating minimum value of  $m$  over all rows that are eligible for processing (have a negative value) where

$$m = \begin{cases} z\%pq & \text{for equality edits} \\ z\%p\%pq & \text{for inequality edits} \end{cases}$$

$z$  is the number of zero elements,  $p$  is the number of positive elements, and  $q$  is the number of negative elements in the row being processed. The value for  $m$  is the maximum number columns that the processed matrix  $\bar{Y}$  can have when processing a particular row of  $Y$ . A row is randomly selected if multiple rows have the minimum value of  $m$ .

Each column of the lower submatrix,  $L$ , contains the correction vectors and the value of  $\delta$ ,  $(y; z; \delta)$ . After each iteration, the columns are re-scaled so that the last entry, corresponding to  $\delta$ , in each column is equal to 1. The generalized cardinality is calculated for each vector  $(y-z)$  listed in the matrix  $\bar{Y}$ . Next, it is checked if a vertex of the matrix associated with  $C_E$  has been generated. A vertex is generated when all elements of a column are non-negative and  $\delta=1$ . The minimum generalized cardinality,  $C_{\min}$ , is calculated for each of the vertices heretofore generated. Any column with generalized cardinality greater than  $C_{\min}$  is deleted since the generalized cardinality associated with a column can never decrease in value (Rubin, 1975). A column corresponding to a vertex with generalized cardinality equal to  $C_{\min}$  is retained only if the pattern of corrections is different from the pattern of corrections of the vertices retained. If a column has generalized cardinality equal to  $C_{\min}$ , but it does not correspond to a vertex (some entries in the column are negative), then it is retained since it can eventually generate a vertex with generalized cardinality equal to  $C_{\min}$ .

## IMPUTATION

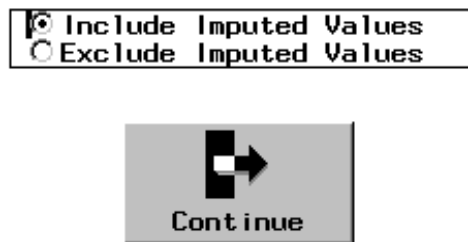
Once data records have been error localized, the values identified to be changed must be imputed such that the imputed values in conjunction with the original values will satisfy the edits. There are several options leading up to the actual imputation of values. The imputation module is selected by clicking on the “Go To Imputation” icon in the output of the error localization module shown in Figure 11.





The “Variable Selection List” listbox contains those variable names with at least one value identified in the error localization module as needing imputation. Clicking on a variable name in this listbox moves the variable name to the “Variable Imputation Order” listbox. Once in the “Variable Imputation Order” listbox, the variable name can be de-selected by clicking on the variable name. This moves the variable name back to the “Variable Selection List” listbox. The resulting order of the variable names in the “Variable Imputation Order” listbox is the order in which the variables are imputed. The current program requires that all variable names be selected. Clicking on the “Continue” icon displays the screen for the second option displayed in Figure 13.

#### Should Imputed Values Contribute to Averages?



**Figure 13.** Include/Exclude Imputed Values in Averages Screen

The second option allows for the exclusion of imputed values when computing averages involved in the imputation estimators. The default option is to include the imputed values when computing the averages. If imputed values are included in the averages, the imputed values for all records imputed prior to the processing of the current record may be used in computing the averages. Sometimes it may be advantageous to include imputed values, while other times it may not. As an example, if most of the imputed values are from larger units, excluding such units would bias the results towards the low side.

Since including the imputed values in the averages of the imputation estimators is the default, this option is selected in the radio box. The selection to exclude imputed values is made by clicking in the circular region to the left of the “Exclude Imputed Values” option. Note, only one selection can be made when using a radio box. Clicking on the “Continue” icon displays the next screen shown in Figure 14.

The next option allows for the specification of the imputation estimators and the order of application for each variable requiring imputation. The six available imputation estimators are (i denotes the unit, t the time period, x the auxiliary variable, and y the survey variable):

Current Mean - the mean of values in the file being edited.

$$y_{it} \quad \bar{y}_t$$

Current Ratio - an auxiliary variable adjusted by the ratio of the current mean to the mean of an auxiliary variable. Only those records that contribute to both averages are used in calculating the ratio.

$$y_{it} = \frac{\bar{y}_t}{\bar{x}_t} x_{it}$$

Previous Value - the value from a previous file for the same unit.

$$y_{it} = y_{i(t&1)}$$

Previous Mean - the mean of values from a previous file.

$$y_{it} = \bar{y}_{(t&1)}$$

Auxiliary Trend - the previous value for the unit being imputed adjusted by the ratio of a current auxiliary variable to the auxiliary variable from a previous file.


$$y_{it} = \frac{x_{it}}{x_{i(t&1)}} y_{i(t&1)}$$


Difference Trend - the previous value adjusted by the ratio of the current mean to the previous mean of the value being imputed. Only those records that contribute to both averages are used in calculating the ratio.


$$y_{it} = \frac{\bar{y}_t}{\bar{y}_{(t&1)}} y_{i(t&1)}$$


Variable :

Variables	Available Selections	Selected Order
LHGFARM1 LHGFDCST LHGFDLBS LHGFDPUR LHGSLDM1	Current Mean Current Ratio Previous Value Previous Mean Auxiliary Trend Difference Trend	

  
**Submit**

  
**View Averages**

  
**Imputation**

  
**Return**

#### **Figure 14.** Specification of Imputation Estimators Screen

The variable names in the “Variables” listbox are in the same order as specified in the “Variable Imputation Order” listbox in Figure 12. To specify the imputation estimators for the variables, a variable name is selected by clicking on the variable name in the “Variables” listbox. This results in the variable name appearing in the rectangular region to the right of the text “Variable :”. An imputation estimator is selected by clicking on it in the “Available Selections” listbox. This moves the imputation estimator to the “Selected Order” listbox. An imputation estimator can be de-selected by clicking on the imputation estimator in the “Selected Order” listbox. This moves the imputation estimator back to the “Available Selections” listbox. The resulting order of the imputation estimators in the “Selected Order” listbox is the order in which the imputation estimators are applied to the selected variable.

Auxiliary variable information is requested as soon as an imputation estimator requiring auxiliary information is selected. After all imputation estimators for a particular variable have been selected, they can then be submitted by clicking on the “Submit” icon. If an imputation estimator has been selected which requires the value of a variable from a previous file, it is requested when the imputation estimators are submitted. Only one previous file may be used for all imputation estimators. The system will not allow the selection of the file being edited as the historical file. In addition, the historical file must contain the same identification variables selected on the file being edited.

Any or all of the imputation estimators may be selected for the variables requiring imputation. If the first imputation estimator results in a value that will result in the record satisfying all edits, then that value is imputed. Otherwise, the next imputation estimators specified are considered. If none of the selected imputation estimators results in a value that will result in the record satisfying all edits, the set of values such that imputing any value in this set will result in the record satisfying all edits is calculated, and the midpoint of this set is imputed.

This approach to imputation more closely resembles the imputation in SPEER than in the GEIS. In the GEIS, a sequence of imputation estimators for a variable can be specified. It is not checked, however, if the imputed value will result in the record satisfying all edits. The sequence is specified just in case an imputation estimator cannot be used (e.g., auxiliary data may be unavailable). The imputation estimators used in the AGGIES are taken from the GEIS. But, unlike the GEIS, it is guaranteed, as in SPEER, that a record will satisfy all edits after being run through the generalized system, since the midpoint imputation method is used as a last resort. The imputing of the midpoint, as a last resort, was taken from SPEER.

The values of the averages involved in the imputation estimators and the number of observations contributing to these averages may be viewed by imputation estimator by clicking on the “View Averages” icon. This aids in the selection of the imputation estimators. An average with too few observations may be unstable resulting in the elimination a particular imputation estimator.

Once all of the imputation options have been selected, the values can be imputed by clicking on the “Imputation” icon. There are two outputs after imputation takes place. The first output displays, for

each edit/data group, the imputation counts by variable by imputation method, including the midpoint imputation method. The second output displays for those records in which one or more values were imputed, the originally reported record followed by the corresponding imputed record. This output, as well as the output from the error localization module, is helpful in establishing an audit trail. Note that if a data record has all of its values identified to be imputed, no imputation will be performed. If this occurs, the record should be reviewed manually.

## EVALUATION OF THE AGGIES

### EDIT SPECIFICATION

The purpose of this section is to show how the edits for the 1996 Iowa quarterly hog report that were specified in the evaluation of SPEER (Todaro, 1997) can be specified, and in many cases, simplified as linear edits in the AGGIES. The number to the left of each edit is the code assigned by the Hog Edit and Analysis Team (HEAT; Anderson et al., 1996). Following the HEAT edit is the edit as formulated for SPEER and the AGGIES (Refer to Appendix 3 for definitions of the variable names). The edits specified for the AGGIES are the linear edits corresponding to the ratio edits specified in SPEER. These linear edits will be used to evaluate the AGGIES.

**501,520**       $lhgund60 + lhgtol119 + lhgtol179 + lhgov180 + lhoggilt + lhogboar - lhogtotl = 0$

SPEER       $lhgund60 + lhgtol119 + lhgtol179 + lhgov180 + lhoggilt + lhogboar - lhogtotl = 0$

AGGIES       $lhgund60 + lhgtol119 + lhgtol179 + lhgov180 + lhoggilt + lhogboar - lhogtotl = 0$

No changes are made to edits 501 and 520.

**505**       $lhgexp13 \# lhoggilt \quad lhgexp46 \# lhoggilt$

SPEER       $1 \# lhoggilt / lhgexp13 \# 4.3$   
                $1 \# lhoggilt / lhgexp46 \# 4.78$

AGGIES       $lhgexp13 - lhoggilt \# 0$   
                $lhgexp46 - lhoggilt \# 0$

SPEER requires that a ratio edit have both a lower and upper bound. In forming the SPEER ratio edit bounds, the resistant fences method (Thompson and Sigman, 1996) generated upper bounds of 4.3 and 4.78. The above two linear edits for the AGGIES are the ratio edits using only the lower bound of one. Including an upper bound may cause unnecessary edit failures and makes the edit more restrictive than that specified by the Hog Edit Analysis Team.

**508**      Pigs born but no sows farrowed

SPEER       $3 \# lhpggsld / lhgfars13 \# 13.5$

4.73#lhpgsld1/lhogfar1#12.36  
 4.73#lhpgsld2/lhogfar2#12.36  
 4.25#lhpgsld3/lhogfar3#13

where

lhpgsld=lhogpig1+lhogpig2+lhogpig3+lhogsld1+lhogsld2+lhogsld3  
 lhgf13=lhogfar1+lhogfar2+lhogfar3  
 lhpgsld1=lhogpig1+lhogsld1  
 lhpgsld2=lhogpig2+lhogsld2  
 lhpgsld3=lhogpig3+lhogsld3

AGGIES     lhogpig1+lhogsld1-4.73lhogfar1\$0  
              lhogpig1+lhogsld1-12.36lhogfar1#0  
              lhogpig2+lhogsld2-4.73lhogfar2\$0  
              lhogpig2+lhogsld2-12.36lhogfar2#0  
              lhogpig3+lhogsld3-4.25lhogfar3\$0  
              lhogpig3+lhogsld3-13.0lhogfar3#0

The AGGIES identified SPEER edit 3#lhpgsld/lhgfar13#13.5 as being redundant. The remaining edits are equivalent for the two systems.

**510** lhogpig3#lhgund60+lhoggilt+lhogboar

SPEER       0#lhogpig3/lhg60brd#1  
               where

lhg60brd=lhgund60+lhoggilt+lhogboar

AGGIES     lhogpig3-lhgund60-lhogboar-lhoggilt#0

The edit specified in SPEER and the AGGIES are equivalent. However, in forming the ratio lhogpig3/lhg60brd, the survey variables lhgund60, lhoggilt, and lhogboar were added to form the variable lhg60brd. As a result the equality (balance) edit lhg60brd=lhgund60+lhoggilt+lhogboar was required in SPEER. But since lhgund60, lhoggilt and lhogboar are involved in other balance edits (e.g., HEAT edits 501 520), the inclusion of this balance edit violated the simple balance edit restriction, that is, a variable can only be involved in one balance edit. Therefore, this balance edit could not be specified in SPEER. The use of the AGGIES averts this problem by using only survey variables in the edit.

**511**            lhogpig1+lhogpig2+lhogpig3#lhgund60+lhgto119+lhgto179+lhoggilt+lhogboar

SPEER       0#lhgpig13/lhg180br#1.33  
               where

lhgpig13=lhogpig1+lhogpig2+lhogpig3  
 lhg180br=lhgund60+lhgto119+lhgto179+lhoggilt+lhogboar

AGGIES      $\text{lhogpig1} + \text{lhogpig2} + \text{lhogpig3} - \text{lhgund60} - \text{lhgtol119} - \text{lhgtol179} - \text{lhoggilt} - \text{lhogboar} \# 0$

The edit specified in SPEER and the AGGIES are equivalent. However, in forming the ratio  $\text{lhgpig13}/\text{lhg180br}$ , the survey variables  $\text{lhogpig1}$ ,  $\text{lhogpig2}$ , and  $\text{lhogpig3}$  were added to form the variable  $\text{lhgpig13}$  and the survey variables  $\text{lhgund60}$ ,  $\text{lhgtol119}$ ,  $\text{lhgtol179}$ ,  $\text{lhoggilt}$ , and  $\text{lhogboar}$  were added to form the variable  $\text{lhg180br}$ . As a result the balance edits  $\text{lhg60brd} = \text{lhgund60} + \text{lhoggilt} + \text{lhogboar}$  and  $\text{lhg180br} = \text{lhgund60} + \text{lhgtol119} + \text{lhgtol179} + \text{lhoggilt} + \text{lhogboar}$  were required in SPEER. Again, the inclusion of these balance edits violated the simple balance edit restriction. Thus, these balance edits could not be specified in SPEER. The AGGIES does not have this restriction and averts this problem by using only survey variables in the edit.

**518**              $\text{lhogpig1} + \text{lhogpig2} + \text{lhogpig3} \# \text{lhogtotl}$

SPEER          $0 \# \text{lhgpig13} / \text{lhogtotl} \# 1$   
                     where

$\text{lhgpig13} = \text{lhogpig1} + \text{lhogpig2} + \text{lhogpig3}$

AGGIES      $\text{lhogpig1} + \text{lhogpig2} + \text{lhogpig3} - \text{lhogtotl} \# 0$

The edit specified in SPEER and the AGGIES are equivalent. However, as in the HEAT edit 511 the AGGIES uses survey variables  $\text{lhogpig1}$ ,  $\text{lhogpig2}$ , and  $\text{lhogpig3}$  rather than creating the variable  $\text{lhgpig13} = \text{lhogpig1} + \text{lhogpig2} + \text{lhogpig3}$ .

**536**             All items not present:  $\text{lhgfdpur}$ ,  $\text{lhgfdcst}$ ,  $\text{lhgfdlbs}$

SPEER          $0.01 \# \text{lhgfdpur} / \text{lhgfdcst} \# 150$   
                      $0.01 \# \text{lhgfdpur} / \text{lhgfdlbs} \# 150$

AGGIES      $\text{lhgfdpur} - 0.01 \text{lhgfdcst} \$ 0$   
                      $\text{lhgfdpur} - 150 \text{lhgfdcst} \# 0$   
                      $\text{lhgfdpur} - 0.01 \text{lhgfdlbs} \$ 0$   
                      $\text{lhgfdpur} - 150 \text{lhgfdlbs} \# 0$

Since SPEER requires that a ratio edit have both a lower and upper bound, the resistant fences method was used to generate upper bounds (150) for both edits. The AGGIES' linear edits are equivalent to the ratio edits. The upper bounds generated for SPEER by the resistant fences method was retained in the AGGIES for detection of possible key entry errors.

**537**              $25 \# \text{lhgfdlbs} \# 120$

SPEER          $0 \# \text{lhgfdlbs} / \text{dumone} \# 120$   
                     where  $\text{dumone} = 1$

AGGIES      $\text{lhgfdlbs} \leq 25$   
                   $\text{lhgfdlbs} \leq 120$

The resistant fences method generated a lower bound equal to zero for the SPEER edit. This lower bound was used since the resistant fences method was being evaluated along with SPEER (Todaro, 1997). However, the two linear edits formulated in the AGGIES are equivalent to that specified by the HEAT edit 537. This was done for comparison purposes. The ratio edit in SPEER required creating the dummy variable, *dumone*, which always has the value of 1. The AGGIES has no such requirement.

538             $0.20 \leq \text{lhgfdlst} / \text{lhgfdlbs} \leq 2.00$

SPEER         $0.20 \leq \text{lhgfdlst} / \text{lhgfdlbs} \leq 2.00$

AGGIES      $\text{lhgfdlst} - 2\text{lhgfdlbs} \leq 0$   
                   $\text{lhgfdlst} - 0.20\text{lhgfdlbs} \geq 0$

The edits specified in SPEER and the AGGIES are equivalent.

The differences in specifying edits in SPEER and the AGGIES can be summarized as follows. First, SPEER requires, both a lower and upper limit for each ratio edit. The AGGIES can accommodate a ratio edit (by linearizing the ratio edit) with either a lower limit and/or an upper limit. Using a method such as resistant fences to calculate the limits for the ratio edits in SPEER may result in the bounds being too restrictive for one limit when the edit only requires the other limit. This can be avoided, however, by arbitrarily specifying a very small limit for the lower limit when only the upper limit is required, or specifying a very large limit for the upper limit when only the lower limit is required.

Second, since SPEER requires the ratio edit to consist of the ratio of two variables, temporary variables, which are the sum of two or more variables, may need to be formed. These temporary variables may violate the simple balance edit restrictions in SPEER. With the use of the AGGIES, no temporary variables are required since the only restriction is that the edits be of linear form. Additionally, the edits involving a temporary variable may actually be identified in the AGGIES as a redundant edit whereas in SPEER it would not. This was the case for the HEAT edit number 508.

## FORMATION OF EDIT/DATA GROUPS

Note that the HEAT edit 537,  $25 \leq \text{lhgfdlbs} \leq 120$ , will fail for the majority of the data records since most respondents apparently do not purchase feeder pigs and, hence, for these records,  $\text{lhgfdlbs} = 0$ . The intention was that this edit should be invoked only for those data records with  $\text{lhgfdlbs} > 0$ . It was noted by Todaro (1997) that whenever the value of *lhgfdlbs* was in the range of 10 to 15 (thereby failing the edit), the values for all three feeder pig variables were usually assigned zero values. This assignment was made based on the statistician's determination that the feeder pig values were unusable. In order to more closely mimic this practice, two edit/data groups were created in the AGGIES as shown in Table 14.

**Table 14.** Formation of Edit/Data Groups

	Group 1	Group 2
Data Group Condition	lhgfdlbs<25	lhgfdlbs\$25
HEAT Edits Forming Edit Group	501,520 505 508 510 511 lhgfdpur=0 lhgfdlbs=0 lhgfdcst=0	501,520 505 508 510 511 536 537 538

In the first edit/data group, edits will be applied to the data group comprised of those data records with lhgfdlbs<25. The second edit/data group will have edits applied to data records where lhgfdlbs\$25. Note that the HEAT edit 518 is not included in either group, since it will always be redundant if the HEAT edit 511 is included. The three edits, lhgfdpur=0, lhgfdlbs=0, and lhgfdcst=0, zero out the feeder pig variables for those data records having lhgfdlbs<25. Thus, the relationships between the feeder pig variables specified in HEAT edits 536, 537, and 538 do not need to be included in Group 1.



### APPENDIX 3–VARIABLE NAMES

Variable	Definition
LHOGTOTL	Total Hogs & Pigs
LHGUND60	Market Hogs & Pigs under 60 LBS
LHGTO119	Market Hogs & Pigs 60-119 LBS
LHGTO179	Market Hogs & Pigs 120-179 LBS
LHGOV180	Market Hogs & Pigs 180+ LBS
LHOGBOAR	Boars & Young Males for Breeding
LHOGGILT	Sows & Gilts for Breeding
LHGEXP13	Sows expected to farrow in next 3 mo.
LHGEXP46	Sows expected to farrow in 4-6 mo.
LHGFAR13	Sows Farrowed the last 3 mo.
LHGFARM1	Sows Farrowed 3 mo. Ago
LHGFARM2	Sows Farrowed 2 mo. Ago
LHGFARM3	Sows Farrowed 1 mo. Ago
LHGPIG13	Pig crop on hand from last 3 mo.
LHOGPIG1	Pig crop on hand from 3 mo. ago
LHOGPIG2	Pig crop on hand from 2 mo. ago
LHOGPIG3	Pig Crop from last mo.
LHGPGSLD	Pigs sold or slaughtered from last 3 mo.
LHOGLD1	Pigs sold or slaughtered from crop 3 mo. ago
LHOGLD2	Pigs sold or slaughtered from crop 2 mo. ago
LHOGLD3	Pigs sold or slaughtered from last mo. crop
LHGFDLBS	Feeder Pig Lb.
LHGFDLST	Feeder Pig Price
LHGFDLPU	Feeder Pigs Purchased