

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (ii): Generalized software packages for statistical data editing, their evaluation

**OVERVIEW AND EVALUATION OF THE AGGIES AUTOMATED EDIT
AND IMPUTATION SYSTEM**

Submitted by the National Agricultural Statistics Service, U.S.A.¹

Contributed paper

I. INTRODUCTION

1. The National Agricultural Statistics Service (NASS), being under tight time constraints to collect, edit and publish survey and census data, constantly strives to improve the efficiency of the editing process. The use of an automated edit and imputation system can improve the efficiency and objectivity of the editing process. At the very least, the functions performed by an automated edit and imputation system are editing, error localization (i.e., identifying which values to change), and imputation. Several automated edit and imputation systems designed to edit continuous data have been developed internationally (e.g., Cotton, 1993, De Waal and Van de Pol, 1997, Draper and Winkler, 1997). This paper describes the current development of an automated edit and imputation system called the AGGIES (AGricultural Generalized Imputation and Edit System) and preliminary results of its application to NASS data.

II. EDITING AND IMPUTATION AT NASS

2. This section discusses current edit and imputation systems used at NASS. The tools that are currently in the survey editing and imputation process are described in order of implementation. For a more thorough treatment of editing and imputation at NASS, see Pense (1997).

A. Survey editing and imputation process

3. The "Survey Processing" System (SPS), developed using SAS in the 1980's, is a batch process. In addition to performing micro-editing, the SPS provides a facility to print groups of records for determining influential and outlying values.

4. The output of the SPS is a printout of records failing one or more edits. These records are manually reviewed by editors. Information for each record includes a descriptive message of the edit failure and the item codes involved in the failing edit along with their associated values. After reviewing each record, the editor pencils in changes on the printout by specifying the item codes to be changed and their associated values. These changes are then entered into the system by key-entry personnel. Once a sufficient number of records have been changed, they are re-submitted in batch overnight. There is no assurance, however, that an edited record will result in the record satisfying all edits. Rather, the changes made by the editor may cause

¹ Prepared by Todd A. Todaro.

other edit failures. This cyclical process of editors reviewing paper printouts, making changes and submitting the changes in batch can become very time-consuming and tedious.

5. In the 1990s NASS began editing interactively using the Blaise system, developed at Statistics Netherlands. In the Blaise system, if a record fails one or more edits, the editor is notified instantaneously, thus eliminating the cyclical batch process. Therefore, the use of the Blaise system can introduce more flexibility in the editing process, reduce costs and increase productivity.

6. The Interactive Data Analysis System (IDAS), written in SAS, is a macro-editing tool that graphically represents the data, allows the capability to drill-down to micro-level data and expands on the SPS facility to print groups of records for determining influential and outlying values (see Apodaca and Hood, 1996).

7. Note that these three tools – SPS, Blaise and IDAS -- are primarily used for automated editing and not automated imputation.

B. Agricultural census editing and imputation process

8. NASS was recently assigned responsibility of the Agricultural Census, beginning with the 1997 Agricultural Census. Edit and imputation procedures used for 1997 were similar as in the past several censuses. The Census edit system allows for batch-processing as well as interactive-processing (i.e., viewing interactively batch-processed data).

9. The Agricultural Census edit system performs automated editing and automated imputation. This is a necessity since it would be a nearly monumental task to manually error localize and impute for approximately 2.67 million report forms that were keyed. However, unlike an automated edit and imputation system which is generalized in the sense that it can be applied to any number of surveys, the Census edit system is not generalized but is specific to the Agricultural Census.

10. The underlying logic of the system, the complex edit, is based on decision logic tables of contingencies and the resulting actions to be taken (McDaniel, 1978). Thus, an edit may be represented as a sequence of contingencies. The action associated with the last contingency specifies the imputation. This approach to editing and imputation can prove to be a very large and tedious task since the description of an acceptable record must consider all of the possible combinations of the contingencies. In addition to the complex edit, the Census edit system also contains a historic logic module that compares a record in the current census to its values in the previous census. The Census edit system has been refined over time while being successfully applied to edit and impute for the Agricultural Censuses. It contains many modules covering pre-processing (processing prior to the complex edit) to summary.

III. RESEARCH INTO AUTOMATED EDIT AND IMPUTATION SYSTEMS

11. In the past few years, NASS has been researching the possible use of automated edit and imputation systems that are based on the Fellegi-Holt model of editing (Fellegi and Holt, 1976). The concept of having a generalized system performing editing, error localization and imputation is very appealing. In 1997, NASS evaluated the SPEER (Structured Programs for Economic Editing and Referrals) automated edit and imputation system developed by the Bureau of the Census. However, because of edit specification restrictions, the SPEER system was not recommended for implementation. While evaluating the SPEER system, NASS became interested in the Generalized Edit and Imputation System (GEIS) developed at Statistics Canada. The GEIS was more generalized than the SPEER system in that the only restriction on the specification of edits was that they must be linear. As a result, it was decided to develop an automated edit and imputation system using the knowledge gained in the evaluation of the SPEER system and possessing many of the same features of the GEIS. Since NASS's experience in SAS is quite extensive, it was decided to develop the system, coined the AGGIES, in SAS.

IV. FUNCTIONALITY

12. The AGGIES was developed using SAS/AF and SAS/IML, products of the SAS Institute, and is a prototype that is being tested for feasibility of use. It is comprised of a number of modules, each performing a separate function.

A. Edit specification

13. Edits are required to be of linear form: $A_1X_1=b_1$ and/or $A_2X_2\leq b_2$, 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.

14. Specifying edits as linear functions of the variables is somewhat different from the traditional manner of formulating edits. 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).

B. Forming edit/data groups

15. The AGGIES includes a module for applying a set of edits (a subset of all edits specified) 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.

C. Check edits

16. In this module, the edits specified are analyzed by edit/data group using linear programming theory. More specifically, the edits are checked for logical consistency, redundancy, hidden equalities, and determinacy. These conditions are checked by solving a series of linear programs (see Giles, 1988).

17. If logical inconsistency is detected, the output consists of a set of edits that, if removed, will result in a consistent 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.

D. Edit summary

18. This module is the first module 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. 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.

E. Outlier detection

19. 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. The addition of a statistical edit module allows greater flexibility in the editing process.

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

21. 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.

22. 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 the variable lies below the Lower Bound or exceeds the Upper Bound, then that value is considered an outlier. This output of this module displays identification information for those records in which a selected variable's value is outlying and is also involved in at least one failed edit. This allows for the possibility of detecting large operations with inconsistencies among the reported values.

F. Error localization

23. 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 to assign based on some criterion. The criterion used for this system is to change the fewest values per record for a record failing edits using Chernikova's algorithm (see Schiopu-Kratina and Kovar, 1989).

24. 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.

25. 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 of the results when running the system on different occasions can be significantly reduced or even eliminated.

26. Occasionally, there are a few records that can consume a large 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 this can occur 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. If the processing time 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.

27. 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. 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.

G. Imputation

28. Once the 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. Two types of imputation strategies will be employed: donor imputation and imputation estimators. The donor imputation strategy uses the nearest neighbor method and is still under development. The imputation estimators strategy allows for the selection and order of application of six available imputation estimators: Current Mean, Current Ratio, Previous Value, Previous Mean, Auxiliary Trend and Difference Trend. Prior to the selection of the imputation estimators, the order in which the variable values are to be imputed needs to be specified. If none of the selected imputation estimators results in a 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, it is guaranteed that a record will satisfy all edits after being run through the AGGIES since a default imputation method is used as a last resort.

29. 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 default 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.

V. APPLICATION

30. The results obtained from the AGGIES have been compared to the results from the survey editing and imputation process using a subset of data from a Quarterly Hog survey. Since the error localization module randomly selects a solution set when several sets, all being minimal, are identified, the AGGIES was run five times to assess the variability of the results obtained. 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. 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 hogs & pigs' 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 likely be manually reviewed by editors because they may have a large impact on the aggregate statistics. However, since the values obtained by the survey editing process were available, they were used as the values for the three outlying records.

31. Table 1 displays, for a subset of the survey variables, the average expanded totals from running the data set five times through the AGGIES, the expanded totals from the survey editing and imputation process and the difference expressed as a percentage of the expanded totals from the survey editing and imputation process. The data were split into two groups according the value of the feeder pig weight variable (not shown below). The first edit/data group contained 14 edits and 1072 records, of which 52 failed one or more edits. The second edit/data group contained 19 edits and 83 records, of which 8 failed one or more edits. The edits consisted of balance, ratio and linear inequality edits. Each group contained 21 variables.

Table 1. Comparison of Average Expanded Totals

Variable	AGGIES Average	Survey Editing Process	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

32. The relatively large percentage difference for ‘boars & young males for breeding’, 7.44 percent, was attributed to the AGGIES changing the value of this variable 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 ‘boars & young males for breeding’ variable values in the survey editing and imputation process.

VI. CONCLUSIONS

33. Using the AGGIES has several potential advantages for NASS:

- Commodity data editing and imputation are performed by the AGGIES resulting in an edited and imputed data set similar to that currently produced using the survey editing and imputation process, as demonstrated using the Quarterly Hog survey. 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.
- The AGGIES 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 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.

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

- The AGGIES will not perform all editing functions. The system is designed for continuous data. Thus, the editing of completion codes and data adjustment factors must be performed outside of the system.
- A plan as to how the AGGIES could be implemented in NASS’s survey editing and imputation processing to form a complete edit and imputation 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.
- 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.

References

- Apodaca, M. and Hood, R. (1996). "Improving the Quality of Survey Data Through and Interactive Data Analysis System," Proceeding of the Twenty-First Annual SAS Users Group International Conference.
- Cotton, C. (1993). "Functional Description of the Generalized Edit and Imputation System," Statistics Canada Technical Report.
- De Waal, T. and Van de Pol, F. (1997). "A Recipe for Applying CHERRYPI in the Edit Process," Working Paper No. 32, Conference of European Statisticians, Work Session on Statistical Data Editing, United Nations Statistical Commission and Economic Commission for Europe.
- Draper, L. and Winkler, W. (1997). "Balancing and Ratio Editing With the New SPEER System," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 81-89.
- Fellegi, I. and Holt, D. (1976). "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, No. 71, pp. 17-35.
- Giles, P. (1988). "A Model for Generalized Edit and Imputation of Survey Data," The Canadian Journal of Statistics, No. 16, pp.57-73.
- Hidiroglou, M. and Berthelot, J. (1986). "Statistical Editing and Imputation for Periodic Business Surveys," Survey Methodology, No. 12, pp. 73-83.
- McDaniel, H. (1978). An Introduction to Decision Logic Tables. Revised Edition. NY/Princeton: PBI..
- Morabito, J. and Shields, M. (1992). "Generalized Edit and Imputation System Applications User's Guide," Statistics Canada Technical Report.
- Pense, R. (1997). "Editing Strategies at the National Agricultural Statistics Service," Working Paper No. 31, Conference of European Statisticians, Work Session on Statistical Data Editing, United Nations Statistical Commission and Economic Commission for Europe.
- Schiopu-Kratina, I. and Kovar, J. (1989). "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Statistics Canada, Methodology Branch Working Paper No. BSMD-89001E.
- Todaro, T. (1997). "Evaluation of the SPEER Automatic Edit and Imputation System," NASS Research Report RD 97-04, Washington, DC: National Agricultural Statistics Service, United States Department of Agriculture.
- Todaro, T. (1999). "Evaluation of the AGGIES Automated Edit and Imputation System," NASS Research Report RD 99-01, Washington, DC: National Agricultural Statistics Service, United States Department of Agriculture.