

STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 31  
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing  
(Prague, Czech Republic, 14-17 October 1997)

**Editing Strategies at the National Agricultural Statistics Service**

Submitted by the National Agricultural Statistics Service, U.S.A. <sup>1</sup>

---

<sup>1</sup> Prepared by Roberta B. Pense

## **I. SUMMARY**

1. The strategies and tools that an Agency uses for editing data are based on such factors as the organizational structure and culture, the methods of data collection, survey timing and volume of data, and the technology available. The National Agricultural Statistics Service (NASS) constantly tries to integrate generalized editing tools to smoothly and efficiently process survey data. NASS also advocates distribution of work to the extent possible. How these strategies have been implemented has evolved as the technological tools to implement them has evolved. This paper describes some of these tools, points out some of their strengths and weaknesses, and describes the plans for further integration based on the use of data warehousing technology.

## **II. BACKGROUND**

2. To understand why NASS uses some of the strategies that it does, one must understand the environment in which NASS operates.

3. NASS uses multiple modes of data collection for virtually every survey. Computer Assisted Telephone Interviewing (CATI) is the predominant mode for most major probability surveys (usually around 50% of the data are collected by CATI) but face-to-face and mail data collection methods are also used. Face-to-face data collection is used for about 15% of the samples, and is the primary data collection mode for respondents with no known telephone number, previous CATI refusals, and extremely large or influential operations. Data collection by mail makes up about 3% of the responses at the US level, but can be up to 20% in some States. The data collection method will affect the volume and type of errors seen in the data. When multiple modes of data collection are used for a single survey, edit procedures must be robust enough to handle all type of errors.

4. NASS has a somewhat unique organizational structure, which takes advantage of the agricultural knowledge of editors in each geographic area. It also adds some complexity to the editing process. Forty-three field offices collect and edit the data under general guidelines set forth by Headquarters in an Survey Administration Manual. For crop surveys, this also means there are 43 versions of the questionnaire and 43 sets of edit limits on variables such as yield. Both State and US level statistics are published, so that outliers need to be identified at both the State and US level, although handling of them may not be the same. Therefore, NASS has used generalized systems with parameters controlling whether edits are invoked or not.

5. Typically NASS allows for a two week data collection period. More burdensome surveys, such as the Agricultural Resource Management Study, allow about six weeks for data collection. There are usually several concurrent surveys running at the same time. For example the December Hog Survey (US sample size of 16,500) and the December Agricultural Survey (US sample size of 56,000) data are collected at the same time (November 29-December 13). State level estimates are then published about two weeks later - December 27 for hogs, and January 9-10 for the crop data. Therefore, only a limited amount

of time is available for editing, which forces the editing to be more focused.

### **III. EVOLUTION OF STRATEGIES AND TOOLS**

#### **A. 1960's - GENERALIZED SYSTEMS**

6. NASS has always focused on using and developing generalized systems. In the 1960's, NASS developed a "Generalized Edit and Summary" system, in COBOL and FORTRAN. The system served as the edit system for all surveys within NASS. Running on a mainframe, statisticians coded somewhat cryptic parameters, which then created the actual micro-level edit logic. Edits could be invoked for a specific State based on values located in specific columns of the parameter. While this allowed State specific edits, the limit of 999 parameters limited the number of edits, especially the use of the State-specific edits.

7. There are many benefits to this generalized approach. Most prominent is the fact that it reduces program maintenance costs by allowing the central core of the system to be re-used for many surveys. While this central core is developed by programmers, the actual coding of edit logic is assigned to statisticians rather than programmers. The editors are therefore able to respond more quickly to problems that they may see in the operational survey edit. The programmers are also freed up to work on more complex and generalized systems. A byproduct of using one generalized system is that users only need to learn how to use that one system. This reduces training costs for both end users (editors) and those preparing the parameters (statisticians). Since NASS traditionally rotates staff to various positions in order to gain a broad knowledge of the entire survey process, minimizing training costs is important.

#### **B. 1980's - FURTHER GENERALIZATION AND NEED FOR INTEGRATION - BATCH MODE**

8. In the 1980's, the Agency goal to reduce respondent burden and minimize survey costs was addressed by survey integration. Thus, several independent State-controlled non-probability surveys were integrated into one National probability survey. This created the need for additional flexibility in our editing system to accommodate State specific needs and edits. The Generalized Edit's limits on numbers and types of edits became painfully obvious. Therefore, NASS developed the Survey Processing System (SPS), using SAS as the programming language, which is a product of the SAS Institute.

9. Some of the features of the SPS are similar to the Generalized Edit. It is a generalized system, with statisticians writing actual validation checking parameter language (SAS code referred to within the system as Usercode) and entering the values into a Specifications Interpreter Module. The Specifications Interpreter Module contains the descriptive information (metadata) for all the variables contained in the survey for any State and is the core of the system. In addition, the Macro language in SAS was used to increase the re-use of code across multiple surveys. For example the "list adjustment factor" coding is common to

all probability surveys and that code is re-used for each of them. Thus, this system took generalization a step further since not only is the system reusable, but actual statistician coding is reusable.

10. The system also included more functionality than its predecessor. Initially the system only contained a module for the micro-level edit. However over time, the system included several macro-editing tools. One of these tools is called Data Listings, which are prints of selected (positive) data, sorted in ascending or descending order. These prints could be pre-programmed at a National level, or created in an ad-hoc manner by statisticians in the States. Another module was the Potential Outlier Prints. These listings used some historical data to identify large changes from previous reports, and incorporated some graphics to more easily identify outliers. As stated previously the Specifications Interpreter Module, containing the descriptive information for all the variables contained in the survey for any State, was the core of the system since all modules shared this information.

11. The use of the Specifications Interpreter Module also allowed statisticians in the States to share in the edit specification. While National limits could not be changed, more restrictive State limits could be specified by the local editor. For example, values greater than a default National edit limit of 500,000 bushels of soybeans in storage might be flagged. But if a State did not produce many soybeans, the State could set the edit limit at 10,000 bushels instead. Thus geographic location (at least at the State level) could be used to determine edit limits.

12. This system runs in batch mode on a leased mainframe computer. Costs therefore limited the frequency of the edits. Potential Outlier Prints (macro edits) were only run twice during the survey - once when 80 percent of the data had been micro-edited, and then again when all the records had passed all of the micro-edits. In the batch edit system, the editor would have to wait overnight to see if the corrections created other edit problems. This could turn into a vicious cycle at the end of the survey. Therefore, the editing strategy was to identify errors earlier in the process. In the extreme, editors manually made the edit checks that would later be made by the computer to correct the record before it was keyed (since it was harder to fix once it was in the computer). This strategy also led to the same edits being invoked multiple times during the edit process, both in micro and macro edits. This redundancy was clearly inefficient, but a pragmatic solution to avoid last minute crises.

### **C. EARLY 1990's - GENERALIZED SYSTEMS AND INTEGRATION - INTERACTIVE MODE**

13. Beginning in the late 1980's, NASS equipped each State office with a Local Area Networks (LAN). Therefore in the early 1990's, NASS started converting from a mainframe batch editing system to a LAN-based interactive system to reduce costs and give States more flexibility in running their edits. Research conducted in NASS (Pierzchala, 1994) also indicated around a 15-20% reduction in post data collection processing for paper collected data by converting to an interactive system because the editor receives instant feedback as to whether the correction created other problems in the data. Rehandling the questionnaires is

therefore eliminated in most cases. While not totally eliminating the need for a hand edit prior to data entry (because of the need to code “office use” codes), it is certainly reduced since correcting errors in an interactive mode is much easier. However, it will take time to “untrain” staff to no longer do processes that served them well for the past 20 years.

14. For micro-editing, NASS uses the Blaise system, developed by Statistics Netherlands. It is designed to integrate many processing needs including data collection (CATI, CAPI, etc.), interactive editing, survey management, tabulation, and meta data management. While Blaise does not meet all of NASS’s needs for all of these functions, it is used for data collection and the associated survey management, and micro-editing. Currently it is used for 16 surveys (some are annual; some are monthly), totaling around 600,000 records annually. When using the Blaise III version, some macro-edit functions are also included through the Maniplus utility. With this tool, developers (statisticians) can create listings of data sorted by important characteristics or relationships. If the editor wishes to review the record in more detail, they can “point and click” on the record and it is automatically opened in the interactive edit mode. The user can, if desired, make changes, and the micro edits are automatically invoked. When the record is stored, control transfers back to the “data listing”, with the output in the data listing automatically resorted.

15. As with the SPS, statisticians write the code in a modular fashion, called BLOCKS in the Blaise system, so that they can be re-used from survey to survey. A Blaise file similar to the Specifications Interpreter was built to handle State-to-State differences in data validation and editing. The same instrument is used to both collect the data (in CATI) and edit the data (collected in CATI or through paper questionnaires). This insures consistency in edits between the two data collection modes and reduces maintenance. However writing code to edit both paper collected and CATI-collected data for the same survey adds complexity to the logic.

16. Since earlier versions of Blaise did not include the Maniplus module, NASS developed its own interactive macro-editing tool, Interactive Data Analysis System, or IDAS (see Hood, 1995, and Apodaca and Hood, 1996). It uses the SAS/AF and SAS/EIS software with some other customized applications based in SAS. This tool is more powerful than the Maniplus utility. IDAS graphically represents the data, allows the editors or commodity analysts to drill down from summary graphics/statistics to micro-level data, and allows analysis over time. See the last page of this paper for screen examples.

17. While integration was technically possible through the use of integrated systems, the application in terms of individual edits was still in the control of the “edit writers”. In a truly integrated system, an individual error should be identified once and the edit check to identify it should be placed in the processing step where the editor gets the most benefit from the check. This would place the edit at the point closest to the source of the information, and remove the cyclical nature of the editing process. However in many cases the redundant edits from the batch mode were blindly transferred to the interactive mode. To address this problem, teams of subject matter specialists were formed to re-evaluate the specification of

our edit.

18. The first team to be created, the Hog Edit and Analysis Team (see Anderson, et al), addressed each edit individually for the Hog Survey. The edits can be grouped as follows:

- 1) Data present when it should be present (ie appropriate routing through the questionnaire). It is invoked in the interview for CATI records; in the micro edit for paper records.
- 2) Relationships between the data items that are physically required. For example: the sum of the breeding and market hogs on hand must equal the total hogs and pigs on hand; the number of cows milked must be less than or equal to the number of milk cows. They are invoked in the interview for CATI records; in the micro edit for paper records.
- 3) Valid entries. For example, if there are 15 categories for “type of farm” then the entry must match one of those categories. They are invoked in interview for CATI records; in the micro edit for paper records. These errors are also invoked in the micro edit for CATI records (that is, in the office after data collection) because we allow “do not know” or “refusal” for any item during the interview. However, we may not allow “do not know” or “refusal” answers for individual items during the micro edit, depending on the type of imputation being used.
- 4) Edits on administrative data. These are edits involving codes entered by the office staff to identify duplication in sampling or adjust for nonresponse. These are invoked in the micro edit for all records. Input of the values should be the only pre-micro edit of paper collected data. However, editors may manually impute data for very large and influential respondents, and review the questionnaire for notes in this pre-micro edit step.
- 5) Flags indicating an unusual but not impossible situations. Examples of these types of edits are checks on the “pigs per litter”, large amounts of grain in storage, or unusually high or low crop yields, large changes from previous values, etc. These are warning errors, since they do not require the value to be changed. For CATI collected data, these are invoked during the interview while the respondent is present and can resolve any differences. For paper collected reports, these are invoked in macro edits so that the distribution of all the data can be viewed. Values falling just outside the “norm” may be left as is, whereas values falling far away for the “norm” would probably be changed. The “norm” would vary by State and season.
- 6) Values not equal to expected. This usually involves expecting periodicity in the data, but observing static values. For example obtaining the same value as last quarter when most other reports showed an increase or decrease. This is a warning error and invoked in the macro edit when the data are aggregated.

19. Unfortunately Blaise and SAS data formats are not totally compatible, so that the micro and macro edits are not fully integrated. Thus changes made in IDAS are not automatically subject to micro-edits. With the implementation of a Windows version of Blaise, SQL compliance should allow that functionality.

#### **D. AUTOMATE ERROR CORRECTION AS MUCH AS POSSIBLE**

20. NASS continues to look for ways to further reduce the need for human intervention in the editing process. Two approaches could be used: 1) an ad-hoc “expert” system, or 2) a generalized Fellegi-Holt system. Given NASS’s tradition of generalized systems, the Fellegi-Holt approach was preferred. NASS chose to look at the SPEER (Structured Programs for Economic Editing and Referrals) system since the source code is readily accessible and it would be compatible with NASS’s database structure (Sybase). SPEER is an automatic edit and imputation system developed by the Bureau of the Census. It is designed to edit continuous data, with the edits specified as ratio edits or balance edits. There is a restriction in the current version of SPEER that a variable can only be involved in at most one balance edit, however. If a record fails one or more edits, SPEER identifies the minimal subset of variable values to delete and impute so that all edits are satisfied for that record. Thus human intervention in the micro-editing process is minimized. NASS hoped to use this approach on those records that were not identified as “extremely influential”.

21. NASS evaluated SPEER using the quarterly Hog Survey for the paper-collected questionnaires (see Todaro, 1997). The balance edit restriction that a variable could only be involved in at most one balance edit was too restrictive for NASS. While 99% of the economic surveys conducted at the Bureau of the Census are not affected by this restriction, 4 balance edits in NASS’s Hog Survey could not be specified because a variable would be included in more than one balance edit. This meant that SPEER would need to be supplemented both before and after its execution with other micro-edits to account for the unspecified balance edits.

22. NASS discovered that very little post-data entry editing is occurring in the Hog Survey. SPEER changed only 88 values out of 14,964 positive data values (in 1155 records), but only 133 values were changed using the current procedures. Of the 88 values being changed by SPEER, almost half would need to be referred to humans to verify the reasonableness of the change. SPEER could not correct 13 records due to the restrictions on the balance edits; 1 record could not be corrected after iterating through the system 6 times; and SPEER made some undesirably large changes to the values of 21 records. In other cases, SPEER could not impute a value because all record level data required for imputation were missing. However, the current procedures may have required much more staff time to review the errors even though corrections were only made to 133 values. Thus SPEER appeared to be doing an extremely good job of identifying the records truly in error.

23. We therefore are looking into the methods used to set the edit bounds in SPEER and apply that to setting limits in the Blaise micro-edit and IDAS macro-edit. The method used is

an exploratory data analysis outlier detection method, called “resistant fences”. The program used to calculate the limits is a SAS program, written by Thompson and Sigman (1996). The resistant fences rules define outliers as those values greater than the upper bound,  $q75+k*H$ , or less than the lower bound,  $q25-k*H$ , where  $q25$  is the first quartile,  $q75$  is the third quartile,  $H$  is the interquartile range ( $q75-q25$ ), and  $k$  is a constant. When the value of  $k$  is 1.5, the rule is called the inner fence; when  $k$  is 2.0, the middle fence rule; and when  $k$  is 3.0, the outer fence rule. No evaluation of these procedures has been done, but it is hoped these procedures can more fully automate the edit limit specifications.

24. NASS will continue to examine methods to automate the error correction and imputation routines, particularly as NASS takes on the responsibilities for the Census of Agriculture. The Bureau of the Census has had the responsibility for this census in previous years, and continues to do the editing and imputation for the 1997 census on a reimbursable basis. However, NASS will need to address the edit and imputation procedures for approximately 3 million census records in 2002.

#### **E. USE OF HISTORIC DATA TO FOCUS EDIT REVIEW**

25. NASS has begun developing a data warehouse (a database or data repository that is optimized for data retrieval and fast data loading rather than on-line transactional processing) to enhance sampling, allow ad hoc analysis, improve survey management, reduce respondent burden and enhance data quality. This “on-line and integrated” specialized database will contain at least 6 years of survey and administrative data and at least two Census data sets. It will contain data at the reporter level, which will allow us to expand the use of historical data during the interview, office edit, and analysis at the respondent level (both within the same survey series and across surveys). Whether the edits will access the warehouse directly or an intermediate data store will depend on the retrieval speed once a large scale pilot system is developed.

26. NASS has previously used historical data for two surveys. The first use was in the development of a PC-based Statistical Edit System for a weekly Census of Livestock Slaughter Plants (Mazur, 1990). The data consist of daily numbers of cattle, hogs, calves and sheep inspected, as well as the weekly weight totals. This interactive edit is based on a robust estimator, called Tukey's Biweight. Each plant's historical data is used to flag outliers, determine which species are normally inspected, see if a pattern is typically followed (eg. whether or not it slaughters on Saturdays), and impute for missing data. This allowed a "custom" edit for specialty (eg. veal) or very large plants, and freed up time to reconcile data problems.

27. The second example involves a survey to determine the value of agricultural land (Fecso, 1996). Values vary considerably by geographic location. Proximity to large cities, interstate highways or water rights can have an extremely large influence on the land's potential value. Setting State or even county level edit values would not accurately identify the records in error. NASS has developed a data history by segment (sampling unit) that is

used to set the limits by segment once past outliers are removed. It also uses graphical techniques to display the range of values within a segment. Segments with extremely large ranges probably contain errors, and can be further investigated with the drill down features in IDAS.

## **References**

Anderson, C. et al (1996). "Report of the Hog Editing and Analysis Team", unpublished documentation, Washington, DC: National Agricultural Statistics Service, United States Department of Agriculture.

Apodaca, M. and Hood, R. (1996) "Improving the Quality of Survey Data through an Interactive Data Analysis System," *Proceedings of the Twenty-First Annual SAS Users Group International Conference*.

Fecso, R. (1996). "Using an Area Frame Survey to Collect Agricultural Land Values," *Proceedings of the 1996 Conference on Methodological Issues in Official Statistics*, Statistics Sweden.

Hood, R. (1995) "Interactive Data Analysis of Survey Data Using SAS/AF and SAS/EIS Software," *Proceedings of the Twentieth Annual SAS Users Group International Conference*.

Mazur, C. (1990) "A Statistical Edit for Livestock Slaughter Data," NASS Research Report SRB 90-01, Washington, DC: National Agricultural Statistics Service, United States Department of Agriculture.

Pierzchala, M. (1988). "A review of Three Editing and Imputation Systems," NASS Research Report SRB 88-10, Washington, DC: National Agricultural Statistics Service, United States Department of Agriculture.

Pierzchala, M. (1994). "Interactive Editing in Blaise," *Blaise Users Group Newsletter* (March), London: Office of Population Censuses and Surveys.

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

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.

## **Interactive Data Analysis System (IDAS) SCREENS**

The IDAS package affords interaction with survey data on a number of levels. The user can move among these levels and summon additional information at the click of a mouse button. This section shows a variety of screens as seen by the analyst.

