STATISTICAL COMMISSION and Working Paper No. 27
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic III: New techniques and tools for editing imputation

DEVELOPING A STATE-OF-THE-ART EDITING AND IMPUTATION SYSTEM FOR NASS' AGRICULTURAL CENSUSES AND SAMPLE SURVEYS

Submitted by the National Agricultural Statistics Service, United States[1]

Contributed paper

I. INTRODUCTION

1. The responsibility for the five-year census of agriculture was transferred from the U.S. Bureau of the Census to the National Agricultural Statistics Service (NASS) in 1997, only months prior to the mail-out of the 1997 Census of Agriculture. The lack of lead time precluded making major changes to the processing systems for the 1997 census; however, NASS has subsequently begun developing a new edit and imputation system for its future censuses and large sample surveys. The motivation for this ambitious project lies in the concern that systems currently used for its various survey programs differ substantially, are somewhat antiquated and, in some cases, lack the desired integration of the editing, imputation and analysis modules. As a result, the senior management of NASS chartered a Processing Methodology Team (PMT) in September 1999 to specify an edit, imputation and analysis system for use in the 2002 Census of Agriculture and subsequent, large NASS surveys. The statement of purpose expressed in the team's charter was as follows:

> A*The objective of the Processing Methodology Team will be to develop systems and procedures that will result in* ***less manual editing, increased interactive editing and analysis capabilities, and a more streamlined data analysis process. Data quality*** *standards must be maintained and hopefully improved. The Team=s recommendations will be utilized across all appropriate survey operations.*
> *The report will define policy and philosophy for these processes that* ***enhance productivity*** *without compromising data integrity. It will also list the features of a* ***user-friendly, effective system*** *from a programmer=s perspective and a statistician=s perspective."*

2. The team's recommendations were formulated through discussions with NASS' Headquarters and State Statistical Office (SSO) staff and documented in a NASS Staff Report (Processing Methodology Team, 2000) published in February 2000. They were then presented to and approved by NASS senior management in March 2000. The specifications provided in the NASS Staff Report and summarized in this working paper are intended to guide developers in designing and developing the new system.

3. The new system must meet the editing, imputation, and analysis requirements of the Agency for both sample surveys and the Census of Agriculture. In the past, separate systems have been developed

[1] Prepared by Dale Atkinson.

and implemented to accommodate the somewhat diverse processing needs of the census and sample survey programs. The main objective of a new and re-engineered edit, imputation, and analysis system is to simplify and integrate this patchwork of disintegrated functional systems into a single, integrated process that enhances productivity, analytical capabilities, and data quality. This proposed system will address the varying requirements of a census and a sample survey through selectable program modules that provide the functionality needed for the task at hand. To be most effective, the processing system must keep track of current data, historical data, and metadata, and make them available to all functional modules. It must also provide the capability to produce canned reports, as well as any ad hoc analysis that may be of interest to subject matter specialists.

4. A single system approach is recommended to ensure that the census and sample survey results are as comparable as possible, to facilitate data sharing between surveys and censuses, and to reduce system maintenance. There should be a common data architecture that serves the data requirements of the system functions of edit, imputation and analysis.

II. THE SURVEY PROCESSING ENVIRONMENT

5. The strategies and tools used by a survey organization for processing survey data are based on such factors as its organizational structure and culture, the methods of data collection, survey timing, uses of the data, volume of data, and the technology available. To understand why NASS uses some of the strategies that it does, one must understand the environment in which it operates.

6. NASS uses multiple modes of data collection for virtually every survey and census. Computer Assisted Telephone Interviewing (CATI) is the predominant mode for most major probability surveys. An analysis of December 1999 Agricultural Survey and Hog Survey returns shows that telephone collection accounted for 59% of the data, while face-to-face and mail data collection combined for only about 13%. Face-to-face data collection is the primary data collection mode for respondents with no known telephone number, for previous CATI refusals, and for extremely large or influential operations. Data collection by mail makes up about 3% of the responses at the US level in the Agency’s annual survey program, but can be up to 20% in some States. In contrast, mail data collection is the predominant mode for the Census of Agriculture. 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.

7. NASS has a somewhat unique organizational structure, which takes advantage of the agricultural knowledge of editors in each geographic area. The structure provides many benefits in conducting surveys, but it also adds some complexity to the editing process. Forty-three field offices collect, edit and analyze the data under general guidelines set forth by Headquarters in a Survey Administration Manual. There are often state-specific versions of the questionnaire and state-specific edit limits.

8. Typically NASS allows for a two week data collection period for surveys. More burdensome surveys, such as the Agricultural Resource Management Study, allow about six weeks for data collection. The Census of Agriculture allows for three months of intensive data collection, although data collection continues for months thereafter. There are usually several concurrent surveys. For example, data from the December Hog Survey (US sample size of 16,500) and the December Agricultural Survey (US sample size of 56,000) are collected at the same time (29 November 13 December). State level estimates are then published about two to four weeks later, depending on the commodity. As a result, only a limited amount of time is available for editing. This forces it to be more focused.

9. Prior to processing the 1997 Census of Agriculture, very few automated data corrections had been made in NASS surveys. The standard mode of operation had always been to have the computer identify potential and definite errors in the data (referred to in the Agency as non-critical and critical

errors), but have statisticians manually correct them. This paradigm of editing had to change with the much larger volume processing associated with a census. In a sense our success in utilizing automated data correction in the census, has helped overcome our corporate culture feeling we need to touch and hand-correct every questionnaire. As a result, the specifications for the new processing system are for it to auto-correct everything. This approach should provide more time for macro-level data analysis, and allow us to focus our limited manual editing time on impact records.

III. GENERAL DISCUSSION OF EDITING AND IMPUTATION

10. A consideration that affects the type of editing implemented is how the data will be used. Data that are only used in aggregate form may require different levels or types of editing from data that are used on an individual basis for modeling or subsampling. In particular, the Census of Agriculture data may be used by data users in ways not originally considered in the design of the edit. For this reason, it is essential that each census record be made internally consistent.

11. Editing should also monitor the total survey process. To allow editing to serve as a process management tool, the edit must include an audit trail. An audit trail is a method of keeping track of changes to values in a field, while also saving the reason and source for each change. Audit trails are generally begun after the initial contact with the respondent, and provide a mechanism to help identify non-sampling errors. If the same error is made repetitively, then cognitive issues may exist and questionnaire design or interviewer training may need to be changed. To be effective the results of the audit trail must be summarized routinely in the initial stages of the edit, so that corrective actions can be taken early in the data collection process.

12. Editing systems can be generalized or programmed very specifically for a survey application. NASS has traditionally preferred the generalized approach , which provides many benefits. Perhaps the most prominent of these is the fact that it reduces program maintenance costs by allowing the core of the system to be re-used for many surveys. While this core is developed by programmers, the actual coding of edit logic is assigned to statisticians rather than programmers. As a result, the editors are able to respond more quickly to problems that they may see in the operational survey edit. The programmers are freed up to work on maintenance and enhancements to the core system, and to develop new state-of-the-art systems. An advantage of using a single generalized system is that users only need to learn one system. This reduces training costs for both end users (editors/analysts) 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 especially important.

13. One of the potential pit-falls in designing a survey processing system is to overly compartmentalize the editing, imputation and analysis aspects of the processing. This approach has created problems in the past. One of the best illustrations of this was manifested in the creation of the imputation program for the Agricultural Survey Program. Since this module was written as something of an afterthought in the development of the Survey Processing System, it was not tightly integrated with the editing process. As an unfortunate result, the program sometimes imputes data values that would not have been accepted by the edit had they been reported. With the sequential nature of the current processing system problematic imputations bypass the micro-edit and go directly to macro-review (and too often to summary). An example of overly compartmentalizing the micro-edit and analysis (macro-edit) functions was evident in the processing of the 1997 Census of Agriculture. Concern has been expressed that too much micro-editing was done with no early indication of its effect. Earlier macro-editing capability would have given the editors a better feel for the data and allowed earlier correction of problems that only appeared much later when the data were aggregated to the county level. Also the use of a score function in the early phases could have been beneficial in focusing manual review on errant records that had a high likelihood of significant impact on aggregate totals. In order to avoid these problems in the future, the micro-editing,

imputation and analysis (macro-editing) functions need to be concurrent and highly interleaved in the new system.

14. In integrated survey processing all parts of the system are assembled in a coherent manner, with one part of the process automatically giving information to the next. The goals of integrated survey processing include one-time specification of the data, which reduces duplication of effort and the numbers of errors introduced into the system. In a truly integrated system, an individual error should be identified once, with the edit check to identify it placed optimally in the processing sequence. This would place the edit at the point closest to the source of the information, and remove the cyclical nature of the editing process.

15. In addition to integration, efficiencies in editing can be achieved by requiring data review on only a subset of records. Selective editing is a procedure which targets only some of the data items or records for manual review by prioritizing the workload through edit thresholds. The thresholds are typically based on models, either graphical or numerical formula based, that determine the impact on the aggregate estimates of specific item values in individual records. The formula or model used to prioritize the manual review is known as a score function.

16. The actual determination of the items to correct for a record to pass all edits (often referred to as error localization) should be based on the Fellegi-Holt (1976) concept of maintaining as much of the reported data as possible. The new system needs to provide the capability of supplying reliability weights for variables so that the determination of specific variable values to change will be based on the minimal "weighted" number of variables. This will provide subject matter experts some additional control of the automated error correction process, by enabling them to minimize the number of times that variables deemed to be reliable are changed.

17. As commonly defined, imputation is a procedure for entering a value for a specific data item where the response is missing or unusable. There are many different ways to obtain this value. If historical data are available for the respondent, these data (perhaps modeled for trend) would likely provide the best imputations. Model-based imputations using highly correlated auxiliary variables can also be effective. In cases where reliable historical data are not available, "nearest neighbor" donor imputation may be effective. Hot deck donor imputation is most useful in situations where the data are fairly dense, and is therefore generally more effective with a census than with a sample survey.

18. The new system must be designed to effectively deliver both model-based and donor imputations. For this to happen it needs direct, seamless access to the NASS Data Warehouse (Nealon, 2000). The system should also allow the subject matter/editing specialist to specify an ordering of variables, indicating the sequence in which they will be imputed. Variable values imputed earlier in the sequence for an operation should be available for use in the imputation of subsequent variables.

19. Expanding our automated processes of error correction/imputation to include those operations currently handled manually could be done with a processing system that benefitted from expanded access to historical data. The PMT advocates the totally automated editing/imputation approach, as it would reduce manual workload and increase consistency in handling non-response across the SSOs. The latter benefit would result in a more statistically defensible product.

20. One final point that should be stressed in planning for a new imputation system is that all imputed values must be such that they would have passed the edit process had they been reported. The new system must be designed with the edit and imputation functions so tightly integrated that it is impossible to impute numbers that do not pass all edit criteria.

21. “Macro-editing” refers to the process of reviewing the data in the context of other records. When a record has been identified as suspicious, the user should be able to drill to the current data, edit the record interactively and see the effect of the micro-edit changes in the macro-edit. Managing the “refresh” operation in the macro-edit screens will be a key issue in optimizing this activity.

22. The 1997 Agricultural Census Analytical Review is an example of a macro-editing tool used at NASS. Within each county in the U.S., tables were generated that compared the 1997 totals to the 1992 Census of Agriculture totals for approximately 2,400 items. When a total in 1997 was far different from its 1992 counterpart, the editor would look at the records contributing the most to that total and could further review records through research screens. This method is sometimes referred to as the aggregation or top-down method (Grandquist, 1990, United Nations, 1994). The method focuses editing on data values that influence the final results.

23. Work on the NASS’ Interactive Data Analysis System (IDAS) began in 1995 and now encompasses most major sample surveys at NASS. IDAS replaced the printouts of unusual values that were formerly used. IDAS allows the user to view records based a scatterplot, then generate a print of extreme observations by simply clicking on the plot. Hood (1995) and Hood and Apodaca (1996) provide details on the programming of IDAS in SAS. IDAS uses the distribution method of identifying problematic data for further review.

24. Many statistical organizations throughout the world are incorporating advanced macro-editing capabilities into their statistical programs that would be beneficial in our new processing system. The Bureau of Labor Statistics utilizes a system for analyzing data by industry group known as ARIES. This system uses the distribution method and has the capability to produce simultaneous scatter plots of multiple variables. The system also utilizes top-down anomaly maps. The anomaly maps look much like branches of a tree with nodes connecting to sub-categories which make up the main nodes. Color coding and other mnemonic devices are used to focus analysts attention toward industry groups where outliers likely exist. This system allows analysts to screen extremely large amounts of detailed data in a short time. A recent upgrade to the system is being implemented to capture the full benefits of a client-server environment (Esposito, et. al., 1997).

25. Another system that shares many design characteristics with ARIES is the GEAQS system, developed and used by the Energy Information Administration (EIA) (Weir, et.al., 1997). This system utilizes anomaly maps and exploratory data analysis (EDA) techniques to help identify unusual petroleum product prices. ARIES and GEAQS use small multiples, box-whisker plots, interactive scatter plots, color coding, and visualization techniques described by Cleveland (1993) to direct attention to unusual reports. A program known as DEEP (Distributed EDDS Editing Program) utilized at the Federal Reserve for analyzing financial information has many design similarities to ARIES and GEAQS. Other macro-editing programs include the Graphical Macro-Editing Application at Statistics Sweden (Engstrom et.al., 1995) and GRED at the New Zealand Department of Statistics.

IV. USES OF THE MACRO-EDIT

26. The macro-edit has two major functions: i) to inform analysts of suspicious records for further review, and ii) to identify situations where the micro-edit or other survey processes need fine tuning. As a review instrument, a macro-edit should be driven by the desired products of the process. That is, the macro-edit should inform the analyst how the published information will be affected by the records that are reviewed (1999 NASS National Conference). In addition to a graphical presentation of the material the analysis system should allow for an interactive compilation of all breakdowns of the data that will be published for a county (or state). Traffic lighting to direct the analyst’s attention toward suspicious aggregates should be employed. “Distance” from previously published census totals, Agricultural Statistics Board estimates, and ancillary administrative totals could be used to highlight cells of

significance. GEAQS uses this concept. The system begins with higher level aggregates and shows via color coding (or "traffic lighting"), aggregates or components that are suspicious. This technique directs the analysts toward aggregates that will "yield" value in the review process. The bottom of these nodes should be a complete replication of the fragment of the particular table that would appear in the state's State and County census publication.

27. In addition to comparing to other published numbers, the macro-edit should inform the analyst about how the micro-edit and imputation algorithms performed. Given the complexity of the edit and imputation systems that will be necessary, analysts will need to be able to effectively "see" what the edit did to the reported data. The analytic screens should also be available early in the processing to serve as a diagnostic tool to fine tune the micro-edit processing. An effective presentation may be to provide a scatterplot of the collected versus the imputed and edited data. Exploratory data analysis techniques of transforming the data may be necessary to produce informative plots (e.g., taking logarithms of the data so that individual points can be distinguished).

V. PROPOSED FEATURES AND DESIRED FUNCTIONALITY

i) Design a system with the necessary scalability to accommodate growth in the number of users and the amount of data.
ii) Ensure that the system is highly integrated, providing a common user interface for all modules of the system to reduce training needs. Provide on-line help and on-line Atutorials.@
iii) Provide seamless access to all current and historical data (including individual reports, survey indications, census aggregates, and official estimates) for detecting and correcting errors, ad-hoc queries and analysis. Provide the ability to drill-down and bubble-up between detail records and aggregate data.
iv) Provide a transactionally efficient, interactive system for processing current data (i.e., capable of updating current data and redisplaying results in real time).
v) Provide on-screen indications as to whether viewed data are reported or imputed. Utilize color and/or symbols consistently throughout the system to denote data source. Be able to recall original pre-edit data at any point in the system.
vi) Provide an audit trail to keep track of changes to data values throughout the system, while saving the reason and source of each change.
vii) Provide an edit monitoring system to track the number of failed or high score records remaining to be manually reviewed.
viii) Provide maximum flexibility in imputation methods.
ix) Ensure that imputed data pass all edits.
x) Provide macro-review capabilities at multiple levels of aggregation and the capability for ad hoc queries and displays.
xi) Provide the capability to reveal multivariate relationships within the data using standard exploratory data analysis procedures.

VI. RECOMMENDATIONS

i) Have the computer automatically correct everything with imputation at the micro-level (i.e., eliminate the requirement for manual review). Allow for manual overrides of the computer actions, but prevent overrides that violate consistency requirements.
ii) To the extent possible, use Fellegi-Holt methodology in the new system. Most of the desirable micro-editing features discussed in this document have already been incorporated in the AGGIES prototype. Therefore, AGGIES should be considered a core portion of the new system.
iii) Use the NASS data warehouse as the primary repository of historic data and ensure that it is directly and seamlessly accessible by all modules of the new system.

iv) *Design the system with tracking and diagnostic capabilities to enable the monitoring of the effect of editing and imputation on total survey quality, and provide error summaries to monitor other aspects of the survey. Develop analytics for a quality assurance program to ensure edited/imputed data are trusted.*
v) *Incorporate a score function to prioritize manual review.*
vi) *Provide universal access to data and program execution within the Agency.*
vii) *Ensure that the system is integrated into the Agency=s overall information technology architecture.*
viii) *Make the system generalized enough to work over the entire scope of the Agency=s survey and census programs.*
ix) *Provide the facility for users to enter and access comments at any point in the system.*
x) *Present as much information as possible on each screen of the system and provide on-screen help for system navigation.*
xi) *Consider the use of browser and Java programming technology to assist in integrating parts of the system across software, hardware, and functions.*
xii) *Designate a developmental team to take this report, develop detailed specifications and begin programming the system.*

VI. IMPLEMENTATION

28. The development of the system should be user focused to ensure the ultimate delivery of a user friendly product. The best way to ensure that it is, is to solicit early input from a broad cross section of users. Including users on the developmental team will also help in this regard. Emphasize communications and cooperation across organizational units to facilitate system improvements.

29. Key modules of the system should be prototyped and populated with demonstration data (using previous census data to the extent possible) for wide-scale review. Show the system's value with concrete applications. Plans should be made to volume test the system with the 2001 Annual Agricultural Survey.

30. In order to gain acceptance and trust and to provide an opportunity for feedback from end users, we recommend a video tape be developed as a training guide for the test modules. This should be delivered to SSOs with the prototype so that key design concepts can be conveyed and evaluated by the largest possible number of users. This iteration will provide invaluable feedback to "mature" the system before it is needed in production. Provide training that coincides with the release of each piece of the system.

REFERENCES

Cleveland, W. S. Visualizing Data, *Hobart Press*, Summit, New Jersey, 1993.

Esposito, R. Lin and Tidemann (1997). AThe ARIES Review System in the BLS Current Employment Statistics Program,@ *Statistical Data Editing Methods and Techniques*, Volume 2 UN/ECE. pp. 84-92.

Engstrom, P. and Angsved, C. AA Graphical Macro-Editing Application@ *Statistical Data Editing Methods and Techniques*, Volume 2 UN/ECE, pp. 92-95.

Federal Committee on Statistical Methodology, *Statistical Policy Working Paper 25*, Data Editing Workshop and Esposition, December 1996.

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association", March 1976, Volume 71, No. 353, 17-35.

Granquist, L. (1990), AA Review of Some Macro-Editing Methods for Rationalizing the Editing Process@ *Proceedings of the Statistics Canada Symposium*, Ottawa: Statistics Canada, pp. 225-234.

Hood, R. (1995) AInteractive Analysis of Survey Data Using SAS AF and SAS/EIS Software,@ *Proceeding of the Twentieth Annual SAS Users Group International Conference*, pp. 760-765.

Hood, R. and Apodaca, M. (1996) AImproving the Quality of Survey Data Through an Interactive Data Analysis System,@ *Proceedings of the Twentieth-First Annual SAS Users Group International Conference.*

Nealon, J. AImproving Our Agricultural Statistics Program Through Easy And Fast Access By Employees To Previous Survey And Census Data@, *NASS Staff Report*, January 2000.

Proceedings of the 1999 NASS National Conference (unpublished).

Processing Methodology Team, ADeveloping a State of the Art Editing, Imputation and Analysis System for the 2002 Agricultural Census and Beyond@, NASS Staff Report, February 2000.

Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology, AData Editing in Federal Statistical Agencies@, *Statistical Policy Working Paper 18, Statistical Policy Office, Office of Information and Regulatory Affairs, OMB*, 1990.

Todaro, Todd A. AEvaluation of the AGGIES Automated Edit and Imputation System@, *NASS Research Report* RD-99-01.

UN/ECE Work Session on Statistical Data Editing, Working Paper No. 2, (Rome, Italy, 2-4 June 1999). ADRAFT GLOSSARY OF TERMS USED IN DATA EDITING" Coordinated by William E. Winkler, U.S. Bureau of the Census.

UN/ECE Work Session on Statistical Data Editing, Working Paper No. 31, (Prague, Czech Republic, 14-17 October 1997). AEDITING STRATEGIES AT THE NATIONAL AGRICULTURAL STATISTICS SERVICE" Prepared by Roberta Pense, National Agricultural Statistics Service.

Weir, P. and Emery, R. AThe Graphical Editing Analysis Query System@ *Statistical Data Editing Methods and Techniques*, Volume. 2 UN/Economic Commission for Europe, pp. 96-104.