Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

GENERALIZED CENSUS AND SURVEY PROCESSING SYSTEM AT THE NATIONAL AGRICULTURAL STATISTICS SERVICE

Carol C. House
National Agricultural Statistics Service, U.S. Department of Agriculture, Fairfax, Virginia 22030-1504

KEYWORDS: generalized systems, editing, survey processing

1. OVERVIEW OF SYSTEM

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture is currently constructing an integrated processing system for the agricultural census and surveys. It will be used operationally beginning January 2003. The system will include modules for record management, record level editing, imputation, micro level analysis, weighting, macro level analysis, summarization, tabulation, disclosure review and cell suppression.

· The record level editing module will support both the standard "if-then" editing formulation and Fellegi-Holt methodology.

· Both the micro and macro level analysis modules will tilt heavily toward graphical analysis. Scatter plots, box-plots and frequency bar charts of various types will be provided. Charts and graphs will be interactively linked to tables and maps. The user will have the option of sub-setting the graph or map by selecting a group of points or by specifying a sub-setting condition. For some plots, the option of additional grouping and/or sub-grouping of a variable(s) through the use of colors and symbols will be available (e.g., by size of farm, type of operation, race, total value of production and other size groups). All graphics will provide drill-down capability to data values and the graphical images of completed questionnaires in order to review and update problematic records. Many of the macro level analysis screens will be designed specifically for the census or individual surveys and will be utilized as appropriate. Tables will be interactive with dynamic sort capabilities.

· The imputation module will support a variety of imputation strategies, with the capability to assign an imputation strategy to each variable separately. In fact, it will be possible to hierarchically define which imputation methodology is employed. Nearest-neighbor donor imputation will play a strong role. The system will leverage the Agency's data warehouse capabilities of providing previously reported survey data.

· The disclosure review module will provide primary and secondary cell suppression based on parameter driven suppression algorithms.

The integrated processing system will be intrinsically linked to important outside systems. Specifically, it will be designed to integrate with the Bureau of the Census National Processing Center's (NPC) mail-out, check-in and tracking system and the NPC's scanning and Intelligent Character Recognition data entry systems. These external systems will be used for mail-out, data entry and tracking of the census through data entry. The processing system will also be linked to NASS's farm register system and its survey management system which will be used for mail-out and tracking of most surveys through data entry. Finally, the system will be linked to the NASS data warehouse which houses previously reported data from the census and surveys. These data will be used for imputation and data analysis and will be a final repository of cleaned data.

The system will be built utilizing a variety of hardware and software platforms. It will utilize Oracle, Sybase and Redbrick databases. Much of the component programs will be written in SAS. It will have various modules running on Windows workstations, an IBM mainframe and Unix mini-computers.

2. GOALS AND OBJECTIVES OF THE GENERALIZED SYSTEM

In 1997 the responsibility for the census of agriculture was transferred from the U.S. Bureau of the Census (BOC) to NASS, providing an opportunity for NASS to improve both the census and

its ongoing survey program through effective integration of the two. The timing of the transfer, however, severely limited the changes NASS could make for the 1997 Census of Agriculture. Much of the data collection, data capture and editing was contracted out to the NPC in Jeffersonville, Indiana. Analysis, tabulation and disclosure review were performed using existing BOC systems. For its ongoing survey program, NASS continued to utilize a SAS based generalized edit and summary system developed initially in the late 1980's, incorporating a variety of enhancements in subsequent years.

NASS has targeted a complete reengineering of the its survey processing systems to effect a proper integration of the census program and NASS' traditional survey program. The scope of this reengineering process is to evaluate the component pieces of both the census processing systems and the survey processing systems, to keep concepts and systems deemed to be effective in an integrated system, and to design and develop new components as appropriate. This system will be available by January 2003 for processing the 2002 Census of Agriculture. Individual surveys will be migrated to the integrated system systematically following the census.

Our guiding principles in developing the new system are as follows:

1) *Automate as much as possible, minimizing required manual intervention*
2) *Adopt a "less is more" philosophy to editing*
3) *Identify real data and edit problems as early as possible in the process*
4) *Design a system that works seamlessly across different platforms and subsystems*
5) *Use the best features of existing products in developing the new system*

3. DEVELOPMENT OF FIRST VERSION

To begin the process of integrating programs, NASS took two major steps. The first of these was to create, in late 1998, the Project to Reengineer and Integrate Statistical Methods (PRISM). The team named to manage this project was charged with conducting a comprehensive review of all aspects of the NASS statistical program and recommending any needed changes. One of their recommendations was to re-engineer and integrate the processing systems used by different parts of the program. The second step was a major reorganization of NASS to help align its organizational structure with an integrated program.

In September 1999 the Processing Methodology Team of PRISM was chartered by senior management to specify a new edit, imputation and analysis system for the 2002 Census of Agriculture and subsequent large NASS surveys. A similar but separate team was formed to review and make recommendations concerning disclosure avoidance methodology. These groups, composed of technical managers and leaders, reviewed literature, existing systems and methodology used in NASS and/or other organizations to synthesize the best of what was available into its recommendations for the new system. These teams published their findings and recommendations in internal reports.

Following the acceptance of the concept teams' recommendations, implementation teams were formed for each component module of the new system. These teams consisted of methodologists, end-users and programmers. It was/is the responsibility of each of these teams to move the development of the system from concept to detailed specifications. As detailed specifications are completed for sub-modules, the programmers begin to write code. The entire implementation team is responsible for reviewing the functionality of the beta system and testing the programs as they are developed. The team is also responsible for modifying specifications if necessary during the programming/testing phases.

The leaders of each implementation team came together as a processing oversight team. Their role was/is to establish regular communication between the various implementation teams to assure that the design of individual modules of the processing system remain in accordance with the overall design of the system. After a year under this structure, a program manager was appointed to facilitate decision making.

4. ADVANTAGES AND DISADVANTAGES OF APPROACH

We had no choice but to build a new processing system for the 2002 census. The processing system utilized for the 1997 census, located within the Bureau of the Census, was being dismantled and was not readily portable to other hardware and/or organizations.

The other decision we made about the system was to integrate the processing of the ongoing surveys on the same system as the census. There were two main reasons for this decision. The first was to enhance data quality. Many estimates such as crop yield and animal production are produced by both the census and the surveys. Different processing methodology (edit rules, imputation, etc.) were contributing to level differences in these estimates. It is our intent to reduce these differences by utilizing an integrated processing system. The second reason for our approach is to enhance future efficiency in maintaining and utilizing separate systems.

An integrated system which can accommodate the census of agriculture must be complex. In addition to the census and several complex multiple frame surveys, NASS conducts a number of small, repetitive simple surveys. Our integrated processing system will probably have too much horsepower and overhead to utilize efficiently for smaller surveys. We are intentionally building the integrated system in modules so that it can be used in a "scaled down" version, but we are likely to find that a number of our ongoing simple surveys will never be converted to the integrated system.

5. DEVELOPMENTAL COSTS

Because the system is still under development, we can only project costs at this point. I estimate the system will cost $15 million to develop and implement. Ninety-five percent of the cost will be for staff. I expect that there will be significant ongoing enhancements to the system after its debut, perhaps adding another $5 million to the overall cost. After the system is fully functioning, yearly maintenance will probably range from $2 to $4 million dollars. We expect the system to be utilized for approximately 10 years.

6. PERFORMANCE MEASURES

Our chief performance measure is that the system is ready for operational use on January 1, 2003, performing as outlined in our specifications.

As secondary measures, we expect to observe the following:

· Fewer differences between estimates of the same commodity from the census and surveys.

· A reduction in staff resources spent on the editing phase of processing with a shift of staff resources to the analysis phase of processing.

· Reduction in errors found immediately prior to or following publication.

· Improved confidentiality protection for respondents through consistently applied cell suppression routines.

· Favorable usability reports from internal users of the system.

· Shorter learning curve for analysts in our State Statistical Offices in carrying out their roles in the processing of data.

7. SYSTEM MAINTENANCE

In our organizational structure, we have units assigned to maintain our survey processing system. Different units maintain the system code, databases and user parameters. We expect these same units to maintain the new integrated processing system. In the short term, however, these units will need to continue to maintain the survey systems until the individual surveys are converted to the integrated system. This will require additional staff resources.

8. LESSONS LEARNED

We will be better able to provide the necessary "hindsight" after our system is fully operational. At this point, I have only a few comments. We needed to start the development of this system earlier than we did, but this was not possible. We had to conclude the 1997 census (published in 1999) in order to understand the basic requirements for the new system. Second, we are happy with our team approach to developing the system. However, we did not do an adequate job of isolating team members from their operational responsibilities. This made it difficult for team members to spend adequate time on design activities in the initial stages, exacerbating an already tight time schedule. Finally, we would have appointed a single program manager earlier in the process.