A Generalized Edit and Analysis System for Agricultural Data

Dale Atkinson and Carol House

The transfer of the census of agriculture from the U.S. Bureau of the Census to the National Agricultural Statistics service provided an opportunity for the Agency to improve both the census and its ongoing survey and estimation program through effective integration of the two. This paper addresses the re-engineering of the census processing system into a generalized edit and analysis system for use on the broad spectrum of agricultural data. The paper discusses issues such as edit philosophy, data capture, macro and micro graphical analysis.

KEY WORDS: edit, analysis, integrated, interactive, seamless, modular design.

1 BACKGROUND[1]

In 1997 the responsibility for the quinquennial census of agriculture was transferred from the U.S. Bureau of the Census (BOC) to the National Agricultural Statistics Service (NASS) in the U.S. Department of Agriculture. This fulfilled a goal of NASS to become the national source of all essential statistics related to U.S. agriculture. It also provided an opportunity for the Agency to improve both the census and its ongoing survey and estimation 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. To complete this census NASS formed a Census Division that had primary responsibility for managing the day-to-day operations of the census activities. This Division included former BOC employees who transferred to NASS with the census. Much of the data collection, data capture and editing was contracted out to the BOC's National Processing Center (NPC) in Jeffersonville, Indiana, which had also assumed these functions in prior censuses.

NASS *was* able to make significant changes in some of the census processes. Specifically, the Agency was able to utilize its 45 State Statistical Offices (SSOs) in coordinating census data collection with that of its ongoing survey program. The SSOs also played a key role in the processes from macro-level editing through the final review of the data for publication. In previous censuses these processes had been centralized and the States' data were reviewed sequentially, in a pre-determined order.

[1]This paper was presented at the **Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR), June 5-7, 2001**. Both authors are with the National Agricultural Statistics Service, Research and Development Division. Dale Atkinson is the Chief, Census and Survey Research Branch. Carol House is the Division Director.

By decentralizing the review process, the States' data were reviewed concurrently -- significantly reducing the time from initial data aggregation to publication. This allowed the publication of 1997 census data a year earlier than those of previous censuses.

However, some of the main benefits of NASS acquiring the census of agriculture have yet to be realized. In particular, a proper integration of the census program with NASS' traditional program figures to improve the quality and efficiency of each. These are benefits that Agency management has targeted for 2002 and beyond. To begin the process of integrating the programs NASS took two major steps. The first of these was the creation in late 1998 of 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. The second step was a major structural reorganization of the Agency. This reorganization essentially absorbed the staff and functions of the Census Division, as formed for the 1997 census, into an enhanced survey/census functional structure. The reorganization was designed to increase efficiency and eliminate duplication of effort by integrating census responsibilities throughout the structure.

2 INTRODUCTION

The census processing system needed to be reengineered prior to 2002. With the transfer of census responsibility in 1997, NASS had inherited an aging system that had been used, largely unmodified, since 1982. It was out-of-date technology-wise and, to a lesser extent, methodology-wise. The system was relatively inflexible in that decision logic tables (DLTs) were "hard coded" in Fortran. It was programmed to run on aging DEC VAX machines running the VMS operating system. While manual review and correction could be performed on standard PC screens, some functionality was lost when the system was used with display terminals other than the amber-screened DEC terminals for which it was designed. In general, the record review and correction process at both the micro- and macro-levels involved navigating an often-frustrating combination of function and control keys. The system had served its purpose through the processing of the 1997 census, but it was time for a more up-to-date system.

In September 1999 the Processing Methodology Sub-Team of PRISM was chartered to specify a new edit, imputation and analysis system for the 2002 Census of Agriculture and subsequent, large NASS surveys. This group reviewed editing literature and processing systems used in NASS and other organizations (U.S. Bureau of the Census, 1996 and Weir, 1996) to synthesize the best of what was available into its recommendations for the new system. In February it published its findings and recommendations in an internal Agency research report. The report highlighted the team's guiding principles, as follows:

1) Automate as much as possible, minimizing required manual intervention – Having dealt exclusively with much smaller sample surveys in the past, the NASS culture has been to touch every questionnaire and have statisticians manually specify needed data changes in response to automated edit flags. The sheer volume of data precludes this option for the census and necessitates a system that makes more editing/imputation decisions automatically, without manual intervention

2) Adopt a "less is more" philosophy to editing – There's a tendency in many organizations to over-edit data -- automatically and/or manually. A leaner edit that focuses on critical data problems is less resource intensive and often more effective than a more complex one.

3) Identify real data and edit problems as early as possible -- One of the concerns about the edit used for the 1997 census was that SSO analysts had nothing to review from the highly automated process for several months after editing started. Except for a few who were temporarily detailed to NPC to correct edit failures, SSO statisticians were unable to see the data until they were weighted for nonresponse and aggregated. This was often six months after initial data collection. The delay caused problems that could have been more effectively handled earlier in the process and imposed additional stress on the SSOs by complicating and compressing their data review time

4) Design a system that works seamlessly – While 'seamless' means different things to different people, what is needed is a system in which all the components interrelate smoothly such that the analyst can quickly and easily navigate to any screen and get any auxiliary data needed to identify and resolve a data problem. A system is definitely not seamless if the user has to log into various computer systems separately to obtain needed auxiliary data or run an ad hoc query. Lack of 'seamlessness' was a problem that reduced the effectiveness of the 1997 census processing system.

5) Use the best features of existing products in developing the new system -- By the time the 1997 Census of Agriculture was completely put to rest, the 2002 Census of Agriculture was uncomfortably close at hand. The short developmental time would preclude "re-inventing the wheel." It was imperative that NASS incorporate the best aspects of what had already been done research-wise and developmentally in NASS and other organizations to expedite the process as much as possible.

In view of the above guiding principles the sub-team documented the features it felt the new system should include (Processing Methodology Sub-Team, 2000). Considerable emphasis was placed on minimizing unnecessary review and on the visual display of data. The sub-team discussed display attributes and methodologies that could be used to identify problematic data with high potential impact on published estimates. The 'features' section of their paper

discussed the issue of refreshing the review screens as error corrections are made and stressed the need for the system to help manage the review process (i.e., to identify records that had already been reviewed, through color and/or special characters). The sub-team concluded its paper with the following recommendations:

i) To the extent possible, use Fellegi-Holt methodology in the new system.

ii) Have the computer automatically correct everything with imputation at the micro-level (i.e., eliminate the requirement for manual review).

iii) Utilize the NASS data warehouse as the primary repository of historical data and ensure that it is directly 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. 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, through modular design, to work over the entire scope of the Agency's survey and census programs.

ix) Enable users to enter and access comments anywhere in the system.

x) Present as much pertinent 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.

3 THE SYSTEM DEVELOPMENT

In response to recommendation *xii*, a number of working groups were formed to focus on various aspects of the processing system development. These included groups addressing check-in, data capture, edit specifications, interactive data review (IDR) screens, imputation, analysis, and census coverage evaluation. In order to ensure consistency of decisions across the working groups in assembling the system an oversight and technical decision-making body, the Processing Sub-Team, was formed of the leaders of the individual working groups. This sub-team was charged with considering the overall system flow and ensuring that the individual modules work together effectively. The sub-team

members keep each other informed about the activities of their individual groups, thus ensuring consistency and that required connectivity is addressed. The sub-team also serves as the technical decision-making body for crosscutting decisions that can't be made by the individual working groups. The following sections describe plans for selected modules of the system, the progress made to date and some key issues that the working groups are grappling with.

3.1 Data Capture

As was the case in 1997, NASS will contract the printing, mailing and check-in of questionnaires and the data capture activities to NPC. While all data capture for the 1997 Census of Agriculture was accomplished through key-entry, NASS' early discussions in preparing for 2002 indicated that scanning could be used to capture both an image of the questionnaire for interactive data review and the data itself, through optical/intelligent character recognition (OCR/ICR). Preliminary testing done with the Agency's Retail Seed Survey supported the practicality of using scanning for data capture. Testing of the OCR/ICR process for this survey was conducted at three different confidence levels (65, 75 and 85%). The outcome of this small test was that at 65%, 4% of the characters were questionable; at 75%, 5-7%; and at 85%, 13%.

NASS will utilize scanning with OCR/ICR as the primary mode of data capture for 2002. Current plans are to start with the industry standard confidence level of 85%, but this might be adjusted with further experience in using the system with agricultural census data. Results from the recently completed census of agriculture content test should help fine-tune the process. Questionable returns will be reviewed, with erroneous data re-entered by correct-from-image (CFI) key-entry operators. The scanning process will produce data and image files, which will be sent to the Agency's leased mainframe computers at the National Information Technology Center (NITC) in Kansas City, Missouri for further processing. The data will pass into the editing system and the images will be brought into the interactive data review screens that will be activated from the analysis system to review and correct problematic data.

3.2 Edit

As the edit groups began to meet on a regular basis the magnitude of the task of developing a new editing system became obvious. The machine edit/imputation used for the 1997 census was enormous. It had consisted of 54 sequentially run modules of approximately 50,000 lines of Fortran code, and the sheer volume of the input decision logic tables (DLTs) was staggering. Through 1997, the census questionnaires had changed very little from one census to the next, so the DLTs and Fortran code had required little modification. For 2002, however, an entirely new processing system would be built on a questionnaire that was also undergoing radical changes. Some of the questionnaire changes were necessitated by recent structural changes in agricultural production and marketing, while others were due to the planned use

of OCR/ICR for data capture. In any case, the group members were saddled with the onerous task of working through the mountainous DLTs from 1997 to determine what routines were still applicable and, of these, which should be included in the new system specifications.

One of the key edit issues is reducing manual review without damaging data quality. In processing the 1997 census data, the complex edit corrected all critical errors and the staff at Jeffersonville manually reviewed ALL "warning" errors. The approximate workload and time invested in this activity follows:

- Approximately 1,800,000 records passed through the edit at least once. Of these, 470,000 (26%) were flagged with warning errors. About 200,000 (47%) of the flagged records required updates.

- About 4,000 staff days were spent performing the review in Jeffersonville.

For 2002, the edit review (and analysis) will be performed in NASS' SSOs. Considering the expected staff shortages in 2002 relative to 1997, the above figures would represent an intolerable commitment of staff resources. Furthermore, indications are that this amount of manual review is not altogether needed or (in some cases) desirable. Table 1 shows the relative impact of 1) the automatic (computer) edit changes with no manual review; 2) edit changes with/from manual review and 3) changes made during analytic review. Due to deficiencies in the edit coding, some changes made totally by computer could not be cleanly broken out from those with manual intervention, resulting in an overstatement of the manual edit effect. All changes made during analytic review resulted from human interaction and are considered part of the impact of manual review.

Table 1 shows that the overall effect of the edit/imputation/analysis process was relatively small for most items, especially crop acreages. Considerably larger adjustments are required for both nonresponse and undercoverage. While admittedly these numbers only reflect the impact on high-level aggregates (U.S. level) and the processing can often be more beneficial at lower levels (e.g., county totals), the size of the adjustments still raises questions about the efficacy of the extremely resource-intensive data editing and review process. Such considerations underpinned our two guiding principles of 1) adopting a "less is more" philosophy to editing and 2) automating as much as possible.

3.3 Imputation

Certainly one of the key considerations in moving to an automated system is determining how to impute for missing and erroneous data. The imputation group is currently working through the census questionnaire "question by question" and "section by section" to

Table1: ***Relative Impact of the Editing/Imputation/Analysis Processing of the 1997 Census of Agriculture Data (U.S. Level)***

Characteristic	Net Effect of Automated Edit Changes (%)	Net Effect of Edit Manual Review (%)	Net Effect of Analytic Review (%)	Total Effect (%)	Total Manual Effect (%)
Corn Acres	(0.24)	(3.97)	0.26	(3.94)	(3.71)
Soybean Acres	(0.20)	(2.33)	0.31	(2.22)	(2.02)
Wheat Acres	(0.69)	(4.18)	(0.01)	(4.88)	(4.19)
Cotton Acres	(0.10)	(0.29)	(0.27)	(0.66)	(0.56)
Cranberry Acres	0.13	1.72	(4.04)	(2.18)	(2.32)
No. of Cattle	0.74	4.75	(0.74)	4.74	4.01
No. of Hogs	0.17	(4.23)	(3.92)	(7.98)	(8.15)

determine the most effective routines to use. Nearest-neighbor donor imputation will play a strong role in filling data gaps. The group is currently developing a SAS®-based donor imputation module, which will provide near optimal imputations in certain situations where high quality matching variables are available. The group will be leveraging the Agency's relatively new data warehouse capabilities of providing previously reported survey data. The data warehouse was populated with the 1997 census data and contains the data from most of the Agency's surveys since 1997. As such, it serves as a valuable input into the imputation process, since many of the respondents in the current survey will have responded to one or more prior surveys. The warehouse data can provide direct imputations in some cases and identify items requiring imputation in many others.

A review of the imputation done for the 1997 Census of Agriculture and the current plans for 2002 indicates the vast majority of the imputation performed will be deterministic (e.g., forcing subparts to equal a total). Deterministic imputation could amount to 70-80% of all imputation for the 2002 Census of Agriculture. Nearest neighbor donor imputation will likely account for 10-20%, while direct imputation of historical data, perhaps 5-10%.

3.4 Analysis

3.4.1 General Description. The analysis system is perhaps the module of interest to the broadest audience in NASS. This module will provide the tools and functionality through which analysts in Headquarters and our SSOs will interact with the data. All processes prior to this point are ones with no manual intervention or, in the case of data capture, one in which only a few will touch the data. As one of our senior executives aptly put it, "All this other stuff – data capture, edit and imputation will happen while I'm sleeping. I'm interested in what will go on when my eyes are open." That's analysis!

Because of the broad interest in and the expected large number of users of the

analysis system, the development team has made a special effort to solicit user input into its specification. The working group chartered to design and program this module circulated a hard-copy prototype of the proposed system to staff throughout the Agency early this year. This exercise resulted in very useful feedback from potential users. The feedback received has been subsequently worked into the module specifications.

3.4.2 Micro-Analysis. After the data have been processed through the edit and imputation steps, during which essentially all critical errors have been computer corrected, they are ready for SSO review in the Analysis System. The first of two analysis phases, micro-analysis, begins immediately. During micro-analysis SSOs will review (and update, if necessary) all records for which imputation was unsuccessful, all records failing consistency checks, and all those with specific items that were flagged for mandatory review. Such records are said to contain critical errors and must be corrected. This work will be done while data collection is ongoing, and will allow ample time for any follow-up deemed necessary. As review time permits the system will also provide the capability to review records that have no critical errors, but may be nonetheless of concern. These would include those identified by the computer as influential or high scoring or with potential problems identified through graphical analytic views. Unlike the 1997 edit, warning errors will NOT be automatically corrected nor require manual intervention

A score function is being developed for 2002 Census of Agriculture to ensure that the records manually reviewed are those that are expected to have a substantial impact on aggregate totals. The quality of county aggregates is of particular concern with the census of agriculture. Therefore, the score function used for 2002 will be one that assigns high scores to records whose current report for selected characteristics represents a large percentage of the previous census' county total for that characteristic.

Micro-level graphics are simply a collection of record level information shown together for all records for specific item(s) of interest. The user will have the option of sub-setting the graph 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, other size groups). Scatter plots, box-plots and frequency bar charts of various types will be provided. All graphics will provide drill-down capability to data values and the IDR screens to review and update problematic records.

Finally, the system will track IDs that have been previously reviewed, compare current values to historic data, allow for canned and ad hoc queries and have a comments feature to document actions. Micro-analysis will also include tables to review previously reported data for non-responding units. This will allow SSOs

to focus nonresponse follow-up efforts on the most "important" records.

3.4.3 Macro-Analysis. The second phase of analysis, macro analysis, begins immediately after preliminary weighting (adjusting for under-coverage and non-response). Macro-analysis uses tables and graphs to review data totals and farm counts by item, county and state. While the macro-analysis tools will retain the key objectives of the analytical review system used for the 1997 census, it will be much more interactive and user-friendly. The focal point of the macro-analysis will be a collection of graphics showing aggregate data at state and county levels. These graphics will include dot plots or bar charts of county rankings with historic comparisons, state maps with counties color-coded by various statistics and scatter plots of current vs. previous data.

The new macro-analysis tool will also be integrated more effectively with the Agency's data warehouse and its associated standard tools for user-defined ad-hoc queries. Graphics or tables will be used to compare current census weighted totals and farm counts against previous census values and other published estimates. There will be a prepared library of database queries, in addition to the ability to build your own. Analysts will drill down to the IDR screens to verify/update records. If micro-analysis is done effectively, the number of issues to be dealt with in this phase will be fewer than in 1997, when no micro-analysis module was available.

The macro-edit can be run as soon as data collection is complete, the last records are run through edit and imputation, and preliminary weights are available. The objective of the macro review will be the same as for 1997. That is, an analyst will be responsible for the complete review of all the state and county totals. According to a state's particular needs and characteristics the SSO's managers can elect to either 1) assign an analyst to a county for the review of all items, 2) have a commodity specialist review items by state and county, or 3) use a combination of both. In any case, every item in each county must be reviewed, and a check-off system will be provided in the analysis system to ensure this is achieved.

4 DEVELOPMENT STATUS

Timelines have been developed for specification and development of the various modules and the groups are working hard to stick to them. Due to a number of factors beyond their control the developmental work started at least a year later than it should have, considering the magnitude of the system overhaul. In spite of the delays and overall staff shortages as compared to what was available for past censuses, the groups have done a fantastic job of moving ahead with the developmental work.

5 ISSUES

One of the key issues in edit development is determining what edits are essential to ensure the integrity of the data without over-editing. This is something that the edit group and

Processing Sub-team have struggled with. The team members represent an interesting blend of cultures. The longer-term, pre-census NASS staff developed within a culture of processing the returns from its sample surveys, where every questionnaire is hand-reviewed and corrected as necessary. While there is a need for some of this extra attention for sample surveys since survey weights can be high, this type of approach is nonetheless prone to manual over-editing. The NASS staff that came over with the census are much more comfortable with automatic editing/imputation and are perhaps overly interested in having the system account for every possible data anomaly. This approach can lead to an excessively complex system that automatically over-edits data.

The combination of these two cultures has resulted in some interesting discussions and decisions relative to the guiding principles of automating as much as possible and adopting a 'less is more' philosophy of editing. Everyone has his or her own pet anecdote indicating a "crucial" situation that a reduced edit would not identify and correct. Such concerns have resulted in some compromises in the editing approach taken for 2002. The new processing system currently being developed will look more like a SAS version of the 1997 edit than the greatly reduced, predominantly error-localization driven system envisioned by the Processing Methodology Sub-Team. The bottom line for 2002 is that there will be 49 DLT edit modules, which will consist of much of the same type of intensive, sequential "if-then" edit conditions that existed in 1997. There are some notable differences in the processes, however. There will be a presence of GEIS-type error-localization (Statistics Canada, 1998) in the new system in addition to the 1997 style editing. Imputation has been moved out of the edit DLTs to a separate module of the system. This module will make strong use of nearest-neighbor donor imputation, enhanced by previously reported data from the Agency's data warehouse. The error-localization presence will help ensure that the imputations will pass all edits. The approach to be used in 2002 will serve as a foothold for the approach (Fellegi-Holt, 1976) initially endorsed by the Processing Methodology Sub-Team. For 2007 there will be a strong push to simplify the edit and increase the error-localization presence or move to the NIM-type approach of Statistics Canada (Bankier, 1999).

Another key issue in assembling the system lies in how much modularity/generality is possible. All efforts currently are, and need to be, directed at having a system in place and tested for the 2002 Census of Agriculture. Due to the tight time frame the developmental team is working with, some compromise on the goal of generality is inevitable. The evolving system is being developed modularly, however; so some retrofitting of generality should be possible.

One of the questions that have yet to be answered is whether runtimes and response times will be adequate? The current plans for the processing system are complex, requiring considerable

cycling through the various sections of the questionnaire. Whether or not the runtimes on batch aspects of the system and response times on the interactive portions will be within workable tolerances will not be fully known until more of the system is built. If the answer in either case turns out to be negative, short- cuts will need to be taken to make the system workable.

Exactly what combination of processing platforms will be used in the final system is another issue that has yet to be fully decided. It will be comprised of some combination of the Agency's leased mainframe, its UNIX boxes and its Windows 98 machines on a Novell wide-area network. Since the system is being written in SAS, which will run on any of the three platforms, the processing platform decision has been delayed up to now. However, in the interest of seamless interoperability it will need to be made soon.

REFERENCES

Bankier, M. (1999), Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses, *Conference of European Statisticians, June 2-4, 1999, Rome, Italy.*

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

Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much is Enough? in L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley and Sons.

Luzi, O. and Pallara, A. (1999), Combining Macroediting and Selective Editing to Detect Influential Observations in a Cross-Sectional Survey Data, *Conference of European Statisticians, June 2-4, 1999, Rome, Italy.*

Processing Methodology Sub-Team (2000), Developing a State of the Art Editing, Imputation and Analysis System for the 2002 Agricultural Census and Beyond, NASS Staff Report, February 2000.

Statistics Canada (1998), Functional Description of the Generalized Edit and Imputation System, Technical report from Statistics Canada.

U.S. Census Bureau (1996), StEPS: Concepts and Overview, Technical report from the U.S. Census Bureau.

Todaro, T.A. (1999), Overview and Evaluation of the AGGIES Automated Edit and Imputation System, *Conference of European Statisticians, June 2-4, 1999, Rome, Italy.*

Weir, P. (1996), Graphical Editing Analysis Query System (GEAQS), *Data Editing Workshop and Exposition, Statistical Policy Working Paper 25*, pp. 126-136, Statistical Policy Office, Office of Management and Budget.