Working Paper 32
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iii): Data editing processes within survey processing

THE DEVELOPMENT AND IMPLEMENTATION OF A NEW PROCESSING SYSTEM FOR THE 2002 CENSUS OF AGRICULTURE

Supporting Paper

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

I. INTRODUCTION

1. The National Agricultural Statistics Service (NASS) assumed responsibility for the quinquennial U.S. census of agriculture from the Bureau of the Census (BOC) in 1997. Since censuses of agriculture in the U.S. are conducted for years ending in '2' and '7' (with initial mail-out at the end of the census year), 1997 was a census year. As such, the timing of the transfer of responsibility severely limited NASS' options for implementing changes for that year's census. As a result, the 1997 Census of Agriculture census was conducted largely as planned by the BOC. NASS' one major enhancement to that census stemmed from its ability to leverage its State Statistical Offices (SSOs) for data editing and follow-up data collection. This use of its infrastructure helped improve data quality and timeliness, resulting in published census results months earlier than those of previous censuses.

2. Upon completion of the 1997 census, NASS started preparing for the 2002 census. One of the Agency's primary objectives in planning for this census was to achieve a better integration of the census work with its ongoing survey and estimation program. The joint use of its infrastructure (e.g., list sampling frames and staff) and enhanced consistency in sampling, data collection, data capture, editing, imputation, estimation and publication procedures would provide an opportunity for better overall Agency products with reduced cost.

3. To effect this integration, NASS embarked on an in-house initiative, termed the Project to Re-Engineer and Integrate Statistical Methods (PRISM). Two aspects of this project were considered essential and served as its cornerstones -- one of these was organizational in nature, while the other was procedural. The organizational cornerstone was optimizing the Agency structure to implement PRISM, while the procedural one was specifying what actually needed to be done for the 2002 census and performing the developmental work to implement the procedures. There was certainly early recognition of the strong dependencies between these two aspects of PRISM, as the organizational changes were designed to facilitate the procedural ones. However, in actual preparation for the 2002 census, they ultimately proved to be intertwined in ways not totally anticipated.

4. The organizational and procedural changes implemented under PRISM will initially be discussed separately in this paper. However, a subsequent discussion of their interaction is pivotal, since a primary

[1] Prepared by Dale Atkinson, dale_atkinson@nass.usda.gov

focus of the paper will be the implications of the organizational changes on the Agency's efforts to implement the procedural ones.

II. ORGANIZATIONAL CHANGES

5. Since October 1986, NASS' headquarters (HQ) has been organized administratively into functionally defined units. This organization is characterized by one unit in the Agency handling a unique subset of the survey activities (e.g., survey management, editing or summarization) for all the surveys it conducts. Prior to the 1986 reorganization, NASS HQ was structured with a program orientation, in which a specific administrative unit was charged with the planning, specification and execution of all aspects of a particular survey program -- with other administrative units doing the same for the Agency's various other survey programs.

6. A primary motivation for the 1986 reorganization was the view of Agency management that a restructure along functional lines would engender improved quality and consistency in handling specific survey functions (e.g., editing) across all surveys. Since the same unit would be responsible for a given survey operation for every survey, each survey would naturally be handled similarly. Also, since a particular unit would only focus on one aspect of the survey operation, that unit could be staffed with the best people for that particular function. The term "centers of excellence" was widely used in describing the expected quality benefits from moving to a functional organization. And, in point of fact, many of the perceived benefits of the functional reorganization *have* been realized. The organizational structure that emerged from this 1986 reorganization has served NASS very well, without creating any significant difficulties in carrying out its programs -- until the arrival of the census, anyway.

7. To absorb the responsibility for the 1997 census on short notice, NASS quickly modified its organizational structure to annex a Census Division to manage all headquarters aspects of the census work. This approach resulted in a hybrid organization in which a program-oriented unit conducted the census, whereas functionally organized units administered the remainder of the Agency's survey and estimation programs. Through the strong efforts of its staff, this census unit conducted the 1997 Census of Agriculture from start to finish very effectively. However, the Agency's management felt that in the longer term, operational efficiency in its overall program would be better served by a more complete integration of the census with its on-going program. Therefore, following the 1997 census, NASS once again reorganized to assimilate more fully the census activity into its overall organizational structure. With this reorganization, the various aspects of the census process were, at least on paper, absorbed into the functional units that handled comparable functions for the Agency's on-going survey program.

8. However, the enormous amount of work that would be required to prepare for the 2002 census resulted in the Agency also creating a number of cross-functional PRISM teams. This approach of cross-functional teams working on a very large, long-term project, in an otherwise functionally structured Agency, resulted in some matrix-management issues. The result was job priority conflicts and problems with collateral duties that would pose some very real challenges for the census preparations. This issue will be discussed further in Section IV of this report.

III. PROCEDURAL CHANGES

9. As planning for the 2002 census began, the need for substantial procedural changes was widely recognized within NASS. Every aspect of the census processing from data capture to publication was viewed as either one that *had* to be reworked, or one that could be significantly improved by some type of re-engineering.

10. Early in the planning process, Agency management decided to utilize scanning technology as part of the data processing for the 2002 census. This technology has been used effectively in a number of statistical organizations throughout the world over the past decade, and the 2002 census seemed like the appropriate opportunity for NASS to implement it. Scanning for image had special appeal for NASS in

its potential to avoid the process of shuttling the census questionnaires from the National Processing Center (NPC) in Jeffersonville, Indiana, the central point for mail-out and mail-back, to the NASS State Statistical Offices (SSOs), where the editing would be done. This cumbersome process of moving large volumes of questionnaires around, and then dealing with them in the SSOs, had occurred for the 1997 census, and the opportunity to avoid repeating it for the 2002 census was widely considered a very appealing thing to do.

11. Less universally accepted was the concept of implementing data capture through the scanning process. There was strong concern about how accurately the optical character recognition (OCR)/intelligent character recognition (ICR) could capture data from the often suspect handwriting of an aging farm population. However, after discussions with other statistical organizations who had used the methodology and some low volume testing of the OCR/ICR process with small-scale surveys and the census content test data, NASS decided to take this additional step. The decision to scan for data capture then spawned another major change, in that the questionnaire had to be reformatted to support OCR/ICR.

12. The core processing system that was utilized for the 1997 census -- featuring micro and macro data editing and hot deck imputation modules -- definitely had to be reworked for the 2002 census. With the transfer of census responsibility, NASS had inherited an aging system that had been used, largely unmodified, since 1982. It was out-of-date technology-wise and, to some 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 1997 census, but it was time for an upgrade.

13. Many of the methodology and technology issues that had to be addressed in the new processing system were interrelated, since the technology available at the time of that system's specification and development had limited what was possible to implement methodology-wise. An excellent example of this was in the micro- and macro-editing systems used. Because of the age of the system used for 1997, these were exclusively text-oriented. A much more graphical approach to identifying problematic data was needed in the new system, especially since the entire editing function would be handled by fewer people for 2002 than had been involved in earlier censuses. For 1997, all the micro-editing was performed at NPC, with only the macro- editing (reviewing aggregate roll-ups of the data) handled in NASS' SSOs. For 2002, *all* editing would be done in the SSOs with, at most, marginally larger SSO staffs.

14. There were also methodological concerns about the legacy system, especially with the imputation procedures for item nonresponse. In 1997, a cell with item nonresponse was imputed with a value from an internal processing system matrix, in which the most recently processed acceptable value for that item was stored. As the next report with an acceptable value for that particular item was processed, the value in the cell of the matrix was replaced with this new value, and it would be used for future imputation needs -- until it was itself replaced. In actual application, this approach resulted in the same value being imputed into too many records for some data items.

15. To address these and many other issues, NASS needed to specify and develop an entirely new processing system, and it needed to be done in a very short period of time. By the time the results of the 1997 census were published in February 1999 and all the census follow-on survey work was completed, the Agency was drawing uncomfortably close to the start of the 2002 census, with a complete re-engineering of a major, complex processing system confronting it. The work on this project didn't start in earnest until fall of 1999 when PRISM was officially launched.

IV. CHALLENGES TO OVERCOME

16. In addition to the short time frame available for the preparations, Agency culture and administrative decisions made in preparing for the 2002 Census of Agriculture -- combined with the Agency's on-going budget battles -- created some interesting challenges along the road to processing the census data. The following paragraphs describe some of the issues that emerged to make the 2002 census preparations an adventure.

17. Agency culture became both a benefit and a challenge in preparing for the 2002 census. On the positive side the NASS staff is accustomed to tackling huge tasks that require significant program changes being made in short periods of time, with meager staff resources. This type of mind-set was certainly both necessary and highly beneficial in preparing for the 2002 census. On the other hand, though, the census of agriculture falls outside the Agency's long-established survey paradigm in a number of ways. Perhaps foremost, it is a much larger survey effort than NASS had ever conducted. This had various implications for editing philosophy in general, and particularly for generating an edit processing system.

18. In terms of editing philosophy, NASS had always dealt exclusively with surveys of a magnitude that every questionnaire could be "touched" and verified to be correct. Furthermore, all records identified to be problematic by its basic survey editing systems had only been flagged as such, requiring an overt action by a statistician to modify the data. No automatic correction of records had been implemented in any of its major survey editing systems. While, this purely manual process was obviously not followed with the 1997 census, there was considerable discomfort expressed by Agency staff at the conclusion of the 1997 census on the "black box" approach of automated data editing and replacement that had occurred there. There was an underlying hope of the Agency's staff that the editing approach for 2002 would feature substantially more "transparency." While, in principle, this sounds like a worthwhile goal, and to some extent it is, in actual practice when taken to extremes for a survey effort in which more than two million records could potentially require review by a relatively small staff, this mindset can create obstacles and delays in finalizing a system design. To some extent it did in our case.

19. Another cultural issue that delayed the developmental effort is the Agency staff's predisposition for "putting out fires." In general, NASS staff members are conditioned to be very due date oriented, and Agency work tends to be of the quick turn-around variety. Also, most NASS surveys are repetitive and conducted on a yearly cycle. Therefore, shortly after one iteration of the survey is completed, plans need to start for the next year's survey. As a result, NASS staff members, already heavily committed to other assignments, tend by necessity to address the most imminent concerns first, while saving other issues until another "due date" approaches and that activity becomes a more immediate "fire." With the same staffs working on both short and longer cycle activities, the shorter cycle ones tend to take precedence. The bottom line was that early-on in the census cycle, without a significant increase in overall staffing levels, census preparations (with the longer 5-year cycle) tended to get shorted in favor of addressing more immediate annual survey program needs.

20. Ultimately, the magnitude of the processing system needs for the census and the struggles with collateral duties that resulted from melding the census into NASS' functional structure created problems for the Agency that significantly delayed the developmental work in preparing for the 2002 census.

V. THE PROCESSING SYSTEM IMPLEMENTED

21. As a result of processing efficiency concerns, as well as delays resulting from some of the issues described above, there were many adjustments made to the processing system plans as developmental work progressed, especially in terms of the editing and imputation processes implemented. Many of the efficiency issues that emerged during the developmental process resulted from the decision to program the system in input/output intensive SAS® -- a decision made primarily because of the pervasive use of SAS® in processing all the Agency's other surveys. To mitigate these concerns, outside consultants were

used extensively to help NASS refine its processing plan and developers fine-tune the code. In particular, the Agency worked closely with SAS® and data base consultants to incorporate as many programming efficiencies as possible into the system. Procedurally, NASS staff visited with staff from Statistics Canada on a number of occasions to gather detailed information on their editing, imputation and scanning procedures to help in finalizing the processing plan.

22. The resulting processing system is a client/server application written in SAS® and sitting on three databases – Oracle, Sybase and Red Brick. The scanning process for data capture and image was contracted out to the BOC's NPC. And since their system was designed to work with an Oracle database, Oracle became a part of the new system and was used to store the images of the questionnaires from the scanning process. Feith software was used to display the questionnaire images, for use in data review. A second database established in the new system, this one in Sybase, was referred to as PRISM-SYS and used to store and manage the administrative information for a questionnaire. Sybase was used for this aspect of the processing system because of its efficiency in handling transactional data. PRISM-SYS would house the data most frequently involved in day-to-day census transactions. Finally the actual data items from the questionnaires were loaded into a PRISM Operational Data Store (PODS) in Red Brick. Red Brick was selected as a result of its strength in quickly extracting stored data, through a Star Schema approach with advanced indexing capability. This is a strength that had previously made Red Brick NASS' database of choice for its data warehouse. The Oracle database and Feith software ran on an IBM RS/6000 R50 server, while PRISM-SYS and PODS ran on the main census processing machine, an IBM Regatta P690 UNIX box with 32 processors and 128 gigabytes of memory.

23. There was a real tension during the specification and developmental processes between user expectations of interactivity and the realities of processing efficiency. NASS analysts expected the level of interactivity they'd been accustomed to in working with data from the Agency's regular surveys, but processing-wise the census turned out to be more than just a "big survey." For some portions of the system, interactivity was simply not feasible. The result was a processing system that consists of both interactive and batch programs. The interactive processes include all the administrative screens, the analysis system, data review, edit and authoring tools, research screens and the incoming telephone call (ITC) tools. The batch processes include the formatting program, which formats the captured data for editing; the wrapper, the program which executes the complex edit; weighting; summary; disclosure avoidance; the management information system (MIS); and various other administrative processes that pass information throughout the system. The system contains nearly 150,000 lines of code in the interactive modules and more than 50,000 lines of code for the batch processes. The following paragraphs summarize the processing system implemented, with special emphasis on the processes directly tied to data editing.

A. SCANNING

24. The scanning for image worked very well. Through the use of Feith software, the image of a questionnaire could be easily retrieved to assist with the data review process, and navigation through the questionnaire using the software capabilities was relatively effortless. The use of the scanned images for data review was a very significant improvement over the 1997 census process of pulling paper questionnaires. Also, the Oracle database running on a UNIX box under AIX proved to be a very stable platform for storing and retrieving the images, with very little downtime throughout the entire data review process.

25. The results from the scanning for data capture, on the other hand, were somewhat mixed. Where respondents had actually entered numbers in the data cells, the results were often surprisingly good, even when the handwriting was less than stellar. However, the scanning created significant problems in situations where additional marks (e.g., explicit decimal points) were entered and especially where respondents crossed out sections, either with lines through the cells or with entire sections "Xd". The crossed out sections often resulted in various combinations of '1s' and '7s' being erroneously captured for some or all of the items in the section. Additional erroneous data captures occurred in situations where respondents wrote comments such as "None" on the form. The various writing styles, angles at

which the text was written, and the size of the text often resulted in unpredictable erroneous values being captured. This variability limited our ability to do mass search and corrections. A substantial amount of staff time was required to clean up the data resulting from these problems!

B. MICRO-EDITING

26. As plans developed for the processing system, it became increasing clear that the editing needs for the census were significantly more extensive than those for other NASS surveys. The census questionnaire is generally longer and contains a greater mix of production agriculture, economics and demographics, with the additional requirement of publishing extensive detail on all these items at the county level. The questionnaire is structured into a series of sections with different types of items (some with cross-sectional dependencies and some not) in each. There is a definite need to prioritize the correction of the various items in the questionnaire. For example, land operated and general land utilization have to be made consistent before acreages of specific crops (in a later section of the questionnaire) can be edited. Also, since different administrative units in NASS are responsible for its crops, livestock, and economic/demographic estimates, there are also different groups of subject matter experts responsible for the results and editing of the various sections. Reasonable internal consistency across all sections of the questionnaire, as well as data quality in aggregate estimates, is required for all the items. The difficulties in meeting these needs were exacerbated by the total number of records to be processed and the need to minimize manual review.

27. As these editing needs were addressed, the system became increasingly lengthy and complex, and modifications were necessary. Early in the developmental process, the initial plans to make extensive use of a Fellegi-Holt (F-H) based statistical edit were greatly reduced. Due largely to the efficiency concerns that had plagued these plans from the start, the "final" system specifications called for a much heavier use of direct IF-THEN-ELSE logic from DLTs, than had been initially envisioned. With the refocus on DLTs, the statistical editing module would be used primarily as a backstop, to fix any problems that had slipped through the DLTs uncorrected. However, during early testing of the editing components, as run-times to process even the DLT modules were substantially longer than expected, further system simplification was deemed necessary. One major simplification implemented was to eliminate the F-H module. Even with the bit role this module had earlier been reduced to, there was an unacceptable amount of system overhead involved in calling it the multiple times that were required for each questionnaire. The multiple calls resulted from the adopted modular edit structure described in the next paragraph.

28. The micro-edit/imputation system developed for the 2002 census consists of 46 modules, most of which address a unique portion of the questionnaire. The questionnaire is edited from front to back, module by module. Each module consists of a DLT, and possibly donor imputation and consistency check routines. The consistency check routines are defined for specific modules where the necessity of rounding and/or ensuring that sub-parts add to a total exists. For each module the basic section-specific DLT is processed first, followed by any donor imputation and consistency check routines designated for that particular DLT module. At the conclusion of a module, the next module's DLT is processed. Once a value is set in a DLT, subsequent modules cannot change it. The penultimate module provides an overall check of consistency across all the DLT modules, and the final module creates summations of the data items for use in analysis and summary. Each record passes through this same sequence of modules, whether it entered the complex edit as part of a batch edit or singly, as a "batch-of-one" from the Data Review tool.

29. As implemented for the 2002 census, the complex edit resulted in more records requiring manual review than had been anticipated. Some of these resulted from systemic problems in the OCR/ICR (discussed previously), and others were simply problems that the system could not fix cleanly. The latter included both records that required correction and those that required review but that could be accepted "as is," if the reviewer found this to be the correct thing to do. The records requiring correction included format, DLT, imputation, and consistency failures. Formatting failures consisted of maximum value failures, where a particular captured value exceeded a pre-established maximum value for that item;

unknown reported counties; and rejected item codes. Other situations requiring review were ones in which the complex edit made the record consistent, but with the internal flags in the edit indicating a low confidence that it had done the right thing. The targeted required review rate from the complex edit for the 2002 census data was less than 10 percent, however due to a number of unforeseen problems in data capture as well as data reporting, the actual rate probably turned out to be well in excess of twice that. This has resulted in the data clean-up process running well behind preplanned timelines.

C. IMPUTATION

30. The imputation process for questionnaire item and section nonresponse underwent many adjustments during system development – most of which were due to system efficiency concerns. From the beginning, plans for imputation were to make maximum use of previously reported data (PRD) to impute for missing data in an otherwise complete questionnaire. (Complete unit nonresponse was handled through weighting.) Since significant numbers of operations on the census mail list had been contacted one or more times over the past several years through the Agency's ongoing survey program, while others had at least reported during the 1997 census, a significant history of these operations was available to be used in imputation. To prepare for the 2002 census, virtually all-applicable survey data collected since 1997 had been stored in the Agency's data warehouse. These PRD, when available, were used by the 2002 census DLTs to impute for missing items through either direct pulls or ratio estimation. This portion of the imputation was implemented with minimal problems and seems to have functioned reasonably well. The larger issues and system changes came about in the fallback procedure where previously reported data were not available, or where the operation seemed to have changed significantly enough that previously reported data were not useful. This fallback procedure was to be the Agency's first use of donor imputation.

31. Early on in the census preparations a specifications team was assembled to hammer out the plans for the donor imputation process, but there was considerable uncertainty on the best approach to take. Donor imputation was new to NASS, and there was not a lot of baseline information to use in making the decisions on how to implement it. Some of the key issues debated were the scope and size of the donor pool, what type of procedure to put in place to limit the size of donor pool as it inevitably grew, what type of distance function to use, and how to evaluate the potential donors for acceptability. Another key issue that surfaced later was the relative efficiency of using skinny vs. wide records. In skinny records, variables are treated as rows, as compared to the standard SAS® "wide record" treatment of them as columns. Indeed, many of the initial decisions made on these issues were changed after initial programming was completed. The changes generally came about as a result of the need to better integrate the imputation SAS® coding with the wrapper (complex edit) code. The lack of familiarity of available programmers in working in a database environment, delays in getting other parts of the system programmed to help guide programming decisions, and the lack of volume test data to test various programming alternatives were of concern.

32. As implemented, the system utilized State-specific donor pools, comprised of all consistent records (i.e., records that had passed the complex edit without error) available at that time, from the particular State of interest and the adjacent agricultural statistics districts (groups of counties) in neighboring States. The wrapper recreated the donor pools on an ongoing basis to keep them current. Matching variables, on which the distance values to determine "nearest neighbors" were calculated, were pre-determined by section of the questionnaire. The values of these were standardized for the computations. Finally, donors were selected by module (section of the questionnaire) from among the 25 nearest neighbors for that module. From these candidates, the first record that passed the linear equations was selected as the donor. The linear equations were designed to ensure consistency of the recipient record after imputation. As the donor pools became large, a sampling procedure was implemented to improve processing efficiency.

33. Not surprisingly there were some problems encountered with our initial use of donor imputation. At times, data imputed from donor records for some modules were not really very consistent with other characteristics of the recipient records. Many of these problems probably could have been avoided or

minimized by research into matching variable and distance function selection, prior to the developmental process. This will be a key area of research for 2007.

34. Perhaps the biggest donor imputation problem encountered to date is the system populating the donor pools with consistent, yet incorrect, data. The problems with these erroneous data, whether they resulted from data capture problems or otherwise, are exacerbated as they are donated second and third hand to other records requiring imputation. To minimize the problems with this type of propagation of erroneous data, the donor pools are regularly recreated rather than updated.

D. ANALYSIS

35. The analysis system was designed to enable analysts to make most effective use of the limited manual review possible with a relatively small number of analysts and about two million records. The developers' goal for the complex edit was to create a subsystem that would deliver consistent records. The developer's complementary goal for the analysis system was to provide a subsystem that would enable SSO and Headquarters analysts to identify impact records from among the edit-processed consistent records that, if left untouched, would have the most negative impact on published data quality. The graphically-oriented tools provided in the analysis subsystem make effective use of data plots and maps with traffic-lighting to attract the analyst's attention to influential, problematic individual data and county aggregates that differ substantially from historical results. A score function, that provides a multivariate assessment of a record's impact on county-level aggregates, is computed and used within the analysis tools. In the case of working from suspect aggregates, the sub-system provides drill-down capability to enable the analyst to identify and update erroneous individual records. Toward this end the analysis tools interface directly with the data review tool.

E. DATA REVIEW

36. The data review tool is the vehicle used to review and update captured data for a given record. While this tool can be entered directly and used to pull up a particular record's ID number, it was designed primarily to be used in conjunction with the analysis tools. Double clicking on a record's representation in analysis, either from a listing or from a point on a chart will bring up the data review tool. Using it in conjunction with the Feith displayed image of the questionnaire enables the analyst to holistically review and update problematic records with full information on what data were recorded on and captured from the questionnaire. All data updated through this instrument are resubmitted through the complex edit as a batch-of-one.

VI. CONCLUSIONS

37. While the developmental process for the new processing system and its implementation with the 2002 Census of Agriculture has been somewhat of a roller coaster ride, it appears that the Agency will emerge successfully from this adventure with a high quality, timely census publication. The developmental process featured a number of false starts, re-do's, and various other bumps along the way, but those involved with the project have to feel a real sense of pride in having ridden out the storm and accomplished so much in so short a time. The processing system developed is much improved over the one used with the 1997 census, and it will serve as a strong foundation for further improvements for 2007. The daring step of forging ahead and revamping essentially everything about the census, during one cycle, has put us well up the learning curve and in excellent position to effect further refinements for 2007. In particular, we gained significant understanding of what works well and what needs to be better accounted for in our overall plan for using OCR/ICR for data capture in the future. It also provided us significant insights on how to operate more effectively in a database environment.

38. In general, most of the problems experienced with the system during the 2002 census were attributable to the lack of both time and an environment for testing. The delays in developing the system resulted in some shortcuts necessarily being taken and in many of the key components of the system

undergoing only minimal stand alone testing -- and virtually no integrated testing. Throughout the developmental process only minimal test data were available. So many changes had been made to the entire system, generating a viable test deck from the 1997 census data would have been a very major undertaking, consuming staff resources that were needed for the developmental work itself. In truth, the first integrated volume test was a live one with the 2002 census data. The recognition of this fact prompted one of the key developers to coin the term "Beta-duction" to describe the processing effort.

39. In conclusion, the system developed and implemented for the 2002 census is in some ways not quite as cutting edge technologically or methodologically as had been initially planned. However, the developmental process is definitely a tribute to what an understaffed, but determined, developmental staff can accomplish by overcoming obstacles of various types in accomplishing a common goal. The effort expended in deploying this first generation of the PRISM processing system has put the Agency in excellent position to incorporate further refinements, possibly including statistical editing, for the 2007 census. While, the culture and structure of NASS might have created some challenges early on in developing the new processing system, ultimately the strong work ethic, perseverance and "can-do" attitude that are so pervasive in the NASS culture enabled its development and implementation to be successful.

References

Atkinson, D. and House, C. (2001), A Generalized Edit and Analysis System for Agricultural Data, *Conference on Agricultural and Environmental Statistical Applications*, Rome, June 5-7, 2001.

Atkinson, D. (2002), Developmental Status of a New Processing System for Agricultural Data, *UNECE Work Session on Statistical Data Editing*, Helsinki, May 27-29,2002.

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.

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.