WP. 47
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2004)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

IMPROVING AN EDIT AND IMPUTATION SYSTEM FOR THE UNITED STATES CENSUS OF AGRICULTURE

Supporting Paper

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

I. INTRODUCTION

1. The National Agricultural Statistics Service's (NASS) first solo experience with the Census of Agriculture in 2002 was a very labor-intensive project for everyone concerned. As with many projects, we had our ups and downs and our setbacks and successes. NASS re-engineered many parts of the overall process to make use of updated technology, to improve methodology and to begin to integrate the newly acquired Census of Agriculture with the existing sample survey programs in the Agency.

2. In processing the 2002 Census of Agriculture data, NASS overcame a lot growing pains and self-induced pains due to the re-engineering of many phases of the 2002 Census of Agriculture. Until 1997, the Census of Agriculture had been the responsibility of the U.S. Bureau of Census (BOC). In 1997, the Census of Agriculture became the sole responsibility of NASS. The most obvious hurdle at the time was how to integrate the Census of Agriculture with the sample survey programs that have existed in NASS for over a hundred years. That task was essentially put off until 2002 and the 1997 Census of Agriculture was conducted in virtually the same way as it had been in 1992 by the U.S. Bureau of Census.

3. The Census of Agriculture differs in many respects from any other NASS program. The NASS survey program is interested in setting commodity estimates whereas the Census of Agriculture emphasizes more demographic types of numbers. Most NASS surveys are conducted monthly, quarterly and annually - while the Census of Agriculture is carried out only once every five years. Finally, most sample surveys are just that - samples of the population of interest (farms) - while the goal of the Census of Agriculture is to receive a form from every farm in the United States. Moreover, the Census of Agriculture is the only NASS questionnaire that the respondent is required by law to complete and return.

4. The Agency decided to re-engineer and update many of the parts of the census process as it was inherited from the BOC. These processes included, but were not limited to the introduction of scanning technology and optical character recognition (OCR) of questionnaires for both data capture and storage and retrieval of the questionnaire image for editing assistance; a questionnaire which had significant changes from the one used in 1997; upgrading of the editing methodology using decision logic tables (DLTs); using a nearest neighbor imputation approach instead of the 'univariate' hot deck used in 1997; and creating an elaborate interactive data analysis system for analysts to review the census data.

[1] Prepared by Jeffrey M. Beranek, (jeff_beranek@nass.usda.gov) and Robert P. McEwen (bob_mcewen@nass.usda.gov).

5. NASS was very adventurous in adopting a 'think big' type of attitude toward the 2002 Census of Agriculture. Some of these ideas were attainable and reasonable, but others were attempted without the necessary resources, planning and testing necessary for the magnitude and number of the proposed changes. Such overreaching meant much more time spent in development of systems than had been anticipated, with the result that testing time was inevitably shortchanged. Still, there were many successes as well. This paper will provide both a technical retrospective of those successes, as well as the failures, in 2002 and a glimpse of the current direction for 2007.

II. EVOLUTION OF THE PROCESSING OF CENSUS OF AGRICULTURE DATA

6. For 2002, NASS contracted the printing, mailing and checking-in of the questionnaires and the data capture activities through the Census Bureau's National Processing Center (NPC) in Jeffersonville, Indiana.

7. In preparation for the 2002 Census of Agriculture and the creation of the Census Mail List (CML), NASS needed to determine the status of 520,000 records on the list frame where the farm operation status was unknown. These records were acquired from other agencies and list sources and deemed to possibly have agricultural activity, but the specific farming activity was unknown. Beginning in the summer of 2002, NASS mailed a Farm Information Survey (FIS) to each of these prospective farming operations to hopefully determine whether or not they should receive a census questionnaire in December of 2002. This activity included two non-response follow-up mailings prior to final creation of the CML. All of these activities were handled by NPC.

8. In December of 2002, NASS mailed 2.85 million list records on the Census Mail List (CML) to farm operators in all fifty states. The completed forms were mailed to NPC and they were scanned using optical character recognition (OCR) software for data capture purposes and intelligent character recognition (ICR) for electronic storage of returned questionnaires. The data from these returned questionnaires were formatted and prepared for the editing process.

9. For 2007, NASS is currently working to increase the list coverage for the census. Again in 2007, there will be names on the list frame where the nature of the farm activity will be unknown. At this time, NASS is trying to decide how to handle these type of records still not resolved by mid-2007.

10. Again, for the 2007 Census of Agriculture, NASS plans to contract the printing, mailing and checking-in of the questionnaires and the data capture activities with the Census Bureau's National Processing Center in Jeffersonville, Indiana.

A. System Changes

11. To facilitate the integration of the Census of Agriculture with the NASS survey programs, the Agency created the Project to Re-Engineer and Integrate Statistical Methods (PRISM). The operational charge of PRISM was to develop and implement procedures needed for the 2002 Census of Agriculture.

12. In preparation for the tremendous task of processing 2.85 million records for the 2002 Census of Agriculture, NASS researched how best to upgrade its computer hardware. Since the PRISM ideology is to integrate survey programs with the agricultural census, a high-end, multi-tasking, multi-processing machine was purchased in 2001. An IBM Regatta P690 UNIX box with 32 processors and 128 gigabytes of memory was purchased to house two of the three main databases used for processing data.

13. In 2002, NASS employed a three-database structure for processing the Census of Agriculture data. An Oracle database was used to house questionnaire images created during scanning. A Sybase database was used for storing administrative data, while a Red Brick database was used as both an archival data warehouse, and as an operational data store. In addition, the Red Brick database was used for transactional

database for updating and as a micro, or record, level edit and as an analytical database for loading, querying and macro, or aggregate, level edits.

14. This design - having two databases store information from the current edited records, treating Red Brick as a transactional database (not really its forte), and running competing processes (editing of live records, and aggregating data, as well as allowing ad hoc research queries to be run against this database) caused major problems during processing. Since batch edits were running almost every hour of the day, any other processes running against the database were competing with the edit for resources. Since two databases were sharing the responsibility of maintaining the 'live' data, two sets of transactions - one for each database - had to be created to update any record, and it often happened that one set of transactions failed to be posted while the other was successfully posted, resulting in a situation where groups of records were 'half updated'. Many problems arose from situations where the administrative data - which had been successfully posted - indicated a record had been edited and was clean, while the value data in Red Brick - which had NOT been successfully updated - indicated unedited, and hence often incomplete and inconsistent - data. In addition, duplicate tables in Sybase and Red Brick that needed to be synchronized in real time. Finally, changes made to data in the Data Review process or by the edit took too long to load to the databases.

15. The system problems were caused by several factors, such as the interactive edit/imputation system design, the data base environment, and the hardware configuration. The slowness of the edit/imputation system significantly hampered the productivity of the data review.

16. In order to solve some of the system related problems encountered during the 2002 Census of Agriculture processing, NASS contracted two separate consultants for the purpose of gaining processing system efficiency. The first was the SAS Institute. One of their recommendations was: "We see no reason to conclude that the current DLT system can be tuned (based on SAS's 2002 and 2004 tuning recommendations) to the point that the speed specification will be achieved.". Since this statement, Computer Specialists at NASS have made a 35-fold improvement in speed by making the existing code more efficient. A second consultant from Destiny Corporation made two recommendations. The first was to redesign and reconfigure the Regatta machine that was used for processing the census data. This recommendation is being implemented. The second recommendation by Destiny requires a complete overhaul of the current system used in NASS. While not deemed realistic or desirable for 2007, it is being considered for the 2012 Census of Agriculture.

17. For 2007, the plan is for Sybase to handle the transactional processing whether it is done in a batch or single record mode. Red Brick will be used for macro level analysis and editing of batches with multiple records. Changes made to data in the Data Review process or by the edit will be temporarily stored in memory and run through the Interactive Edit (IE) before loading to Red Brick. Questionnaires scanned for image will again be kept in an Oracle database on the Regatta, though NASS is evaluating the cost and difficulty in converting this image storage system to Sybase. This restructuring of the databases and the re-engineering of how these databases relate to one another will go a long way to improving the efficiency and reliability of the overall processing system.

18. To rectify the overall testing deficiency, NASS has planned to have all code and systems upgraded and fully integrated by January 1, 2007. During 2007, NASS will conduct full system testing using both 2002 census data and content test data. Content test data is data collected during the 2005 testing of the proposed changes made to the questionnaire for 2007. This full system test can be repeated during the reference year of 2007. It is the intent that when actual 2007 data is processed in 2008, that the system will behave without too many unforeseen problems and that the system will make steady, every day progress towards completion of the 2007 Census of Agriculture.

B. Data Capture

19. For the 2002 Census of Agriculture, NASS employed scanning and optical character recognition (OCR) software for data capture. This process was performed at NPC after the questionnaires were checked-in. These data were then used as the input to the editing portion of the system. The scanning of

questionnaires for both data capture and storage and retrieval of the questionnaire image for editing assistance was a new technology for the 2002 Census of Agriculture.

20. Questionnaire images captured through OCR were available for display during editing and data review. The questionnaire images were stored in a Oracle database on a UNIX box named FEITH. When an analyst needed to consult the farm operator completed questionnaire to verify data or to help analyze/review the data, all they had to do was login to the FEITH server and enter the questionnaire ID and the questionnaire image would appear on their screen. The analyst could then toggle back and forth between the Data Review screens and the questionnaire image to resolve any problems or data inconsistencies.

21. Scanning and capturing data using OCR caused considerable problems during the processing of 2002 Census of Agriculture data. The BOC had conducted tests to rate the reliability of scanning for data capture and decided not to use this approach for capturing data. Statistics Canada employs scanning for data capture, but most of their surveys are population censuses or household type of survey that are very straightforward and give the respondent very little opportunity to make the scanning process fail. These errors consisted of incorrectly scanned data and respondent marks on the questionnaire that were never intended to be reported data. Because the 2002 Census of Agriculture was a mail questionnaire, it was impossible for the most part to control how the questionnaire was completed. Many respondents would draw lines through entire sections of the questionnaire that did not pertain to their farm operation. Check marks or "X's" in certain cells would often stray into neighboring cells; these were interpreted by OCR as legitimate data (usually '1') and captured, to the resulting detriment of data quality. The OCR software was also not able to clearly distinguish the many different variations in handwriting. NASS had never tried to scan for data capture on any of its sample surveys, so trying for the first time on a Census of Agriculture was very short-sighted.

22. Scanning and capturing the image of the returned questionnaire was very well received in the Agency and proved to be a very efficient way for analysts to consult with the reported data to assist with Data Review and Analysis. In previous censuses, the paper questionnaires were returned to NPC in Jeffersonville and retained there; this is the procedure NASS followed in 1997. This meant, of course, that analysts in the NASS state field offices did not have access to the actual form that the operator had filled out. Scanning the form, and storing its image, provided a way for all NASS offices, and headquarters, to have access to this information.

23. Because of the problems encountered when capturing data using OCR software, NASS has decided to abandon this approach for 2007. Data instead will be keyed from the image. This is the same image captured for data review, analysis and storage.

24. During the data collection process, NASS will employ a system of verifying the data keyed from image. Not all forms will be verified, but a certain sample of them will. Verification consists of re-keying a form to prove that it was in deed keyed correctly the first time. Quality control measures will be different for image scanned for storage. During their review, analysts will notice inconsistencies between data in the Analysis System and the image on their screen. This could be due to such factors as: missing pages in the database image, differences in the data, and missing data that should have been keyed.

C. Edit

25. The 2002 Census Edit was based on a series of decision logic tables (DLTs), which provide an easy way for subject-matter experts (typically non-programmers) to encode if-then type logic when developing edit code. This program structure was in place when NASS took over responsibility for the Census of Agriculture, but DLTs are also used by the Department of Defense. In practice, DLTs function as both documentation for the edit code, and pseudo-code for the edit program itself. They are particularly useful in making the content and structure of an edit system more transparent to a non-programmer than would the actual edit code.

26. In previous censuses under the BOC, the DLTs were developed and maintained by subject-matter experts, and the completed tables would be taken to programmers who would write code (FORTRAN) which performed the desired checks and actions, but which was written in the most efficient form, so that the actual edit code may have differed greatly in structure from the DLTs. This had the benefit of making the actual edit code run as efficiently (hence, presumably, as quickly) as possible, but had the disadvantage of introducing an additional step between the subject-matter expert and the edit code. The subject-matter expert could not directly test the actual code. Moreover, if changes were necessary, these changes would have to first be made to the DLTs, then incorporated into the edit code by the programmer. In contrast, in 2002 NASS developed a utility referred to as an 'Authoring System' - more or less a code generator - which in effect removed the middle step of having a programmer turn the DLT into executable code. In 2002 this code took the form of SAS SCL lists, which were parsed and interpreted by NASS-written SAS/AF utilities into the edit code. This meant that the DLT authors were able to test their own code, and more easily and quickly able to test and implement updates if they were needed. Obviously, such generated code is bound to be less efficient than code written by programmers, but NASS deemed the trade off worthwhile.

27. In 2002, as in previous censuses, DLTs were grouped into processing units called modules, where each module edited a set of related data. For example, a given module edited cattle inventory data, while another handled data collected on vegetable acreage. There were 46 of these modules written for the 2002 Census. These DLTs were responsible for checking each questionnaire for consistency and completeness in the reported data. If the data were incomplete, or invalid, the DLT code attempted to fix them by imputing correct values for the missing, or invalid, reported ones. Typically "deterministic imputation" was tried first. This was possible, if, for example, there was one unknown field among a group of fields which must add to a total - effectively this was solving one equation with one unknown. If the fields did not admit of this solution, another source of imputed data was a record's own previous reported data (PRD) from some other NASS survey or surveys. A database of PRD was created prior to processing census forms and contained farm operator reported data from the previous census or from a recent sampled survey.

28. Suppose, for example, an operator reported his total acres in bearing and nonbearing fruit orchards, but not any of the acres for the individual types of fruit. If these individual acres existed as PRD for this record, this data could be used (appropriately adjusted if necessary) to 'fill in' the data record for this operator. Failing this, a hot deck approach using Nearest Neighbor Imputation (NNI) was used. Twenty-eight of the forty-six DLT modules made use of the NNI approach. Once edited and clean, 2002 Census of Agriculture records (with a few exceptions) were placed in state-specific donor pools. The donor pools were searched using variable-specific matching variables and a very simple Euclidean distance measurement was calculated between the record needing imputation (recipient) and all donors in the donor pool. Data from the donor 'closest' to the recipient was used to impute for the required data on the recipient record.

29. Indeed, a major success in developing the 2002 Census of Agriculture edit was the use of the DLTs. The authors (some of whom hadn't worked with DLTs before) were very enthusiastic about the ease with which one could learn to write them, and appreciated how 'transparent' they made the edit code. In addition, the Authoring System contained utilities that would allow the author to create test records, run them through the module code, and examine the output and trace the flow of processing. If errors occurred, or the output was unexpected, the trace allowed the author to quickly discover the source of the error in the DLTs and as quickly correct it. This allowed the authors to test their code at the module level, and incorporate corrections, much more quickly than had been possible in 1997.

30. Testing at the record level, running a test record through the entire set of 46 modules, with NNI incorporated, could not be done in the Authoring System in 2002, as it did not have the capability to mimic NNI. Changes in the DLT_Edit system (the code which runs the parsed DLT code) will make it possible to run a staged record through the entire edit in 2007, and will include the ability to 'stage' donor data for imputation as well. While adding to the complexity of the testing procedures, this will enable us to test the entire system more rigorously before moving on to testing with 'real' data.

31. Every record was processed in a batch the first time through this edit. Ideally, the record would be

edited without needing any more intervention. However, there were many circumstances that might have resulted in the record needing to be reviewed by an analyst. It might have happened that a record needed to have data imputed via NNI from a donor, but a suitable donor could not be found. These 'imputation failures' halted editing, and an analyst had to review the record.

32. Another source of analyst reviews were data that the edit had to change to such an extent that, even though it was now consistent, an analyst review was deemed to be desirable. There were also cases where the edit code could not make a determination as to how to fix the data. These were also marked for analyst review. The original goal for the 2002 Census of Agriculture was to have no more than 10% to 15% of the records 'touched' by analyst in this way; the actual figure was closer to 25%.

33. The utility created for single record review (and possibly reediting) was called the Data Review (DR) system. In it the analyst could view the current state of the record, as well as historic data associated with the record, review flags and correct errors, and resubmit the record to be edited with corrections, or by resetting the record to its original reported data. While this process was referred to as an interactive edit, in design, and in practice it was anything but that. It is more accurately described as a batch edit of one record, since the processes behind it - creating and posting transactions, and making the record available to the edit - were exactly the same as those for multiple-record batches. Nor were these 'batches of one' given higher priority than the other batches, so that a person submitting a record from DR to be reedited might not have seen that record again for several hours - or days.

34. Data review by analysts in Data Review and the Analysis System will be the most beneficial quality control measure for the DLT modules. Analysts will identify which edits fail most frequently and which data relationships may indicate the need for an additional edit or a change to an existing edit.

C.Imputation

35. When designing an imputation strategy, one of the first problems to solve is that of donor pool creation. In 2002, NASS used state-specific donor pools, populated with clean, consistent records from that state and adjacent counties in neighboring states (although certain 'abnormal' records, such as prison farms and research facilities, were excluded). Hence, the donor pool for Florida would have also contained records from Georgia and Alabama. One difficulty was the question of how to 'start' the donor pools. In 1997, imputation was not 'started' until 'enough' records had been edited. In 2002, rather than have two processes - one when editing was just beginning, the other when editing had been ongoing and a sufficient number of current records had been edited and cleaned - NASS decided to try to 'seed' the donor pool with staged records prior to the start of editing, so that the donor pools would be available as a source of imputation data at the very beginning of production processing. There were two difficulties in attempting this, however. One was that the 2002 form was very different from the 1997 form, so that the 1997 data could not very easily be 'remapped' to the 2002 record layout and serve as a source of initial donor records. The second was that at the time we needed to create these 'startup' donor pools we did not yet have a working edit, so even if we had decided to use the 1997 data, we could not run it through an edit to guarantee it was clean and consistent by 2002 standards. Ultimately, we had to fall back on a set of 'staged' clean donor records, created by NASS personnel, and based on NASS survey data and the 1997 Census data. While it was hoped, even in the absence of having actually been edited, that these records were 'clean', many imputation problems which came to light later during processing, and which were ultimately traceable to this starter set of records, proved that this was not the case.

36. The donor pools were SAS data sets that were recreated on a regular basis. As more records were processed, the donor pools were recreated to ensure that the best donors were available for every future imputation needed. Initially the donor pool for any given state consisted of all the available clean records in that state and adjacent counties, and were recreated every day. As production processing progressed, and the donor pools became bigger, they were refreshed at more infrequent intervals - every other day, to once a week, to perhaps once every few weeks. In addition, the donor pools for most states eventually became so large that most were eventually sampled, so that not every clean consistent record for a state ended up as a potential donor.

37. Once sufficiently deep donor pools were created and available, a donor had to be found. Each item on the census questionnaire was identified with a set of matching variables. These matching variables were administrative or reported data items that were closely correlated with the census item needing imputation. For an item needing imputation, an N-dimensional, normalized, Euclidean distance was calculated for the matching variables of the item needing imputation on the recipient record and that of all donors in the pool. After a distance is calculated to each possible donor from the recipient record, the record in the donor pool with the smallest distance was called the nearest neighbor and its data were used to impute for all unanswered items in the recipient record. Calculating all these distance to find a nearest neighbor was computationally intensive.

38. After the donor data was written to the record needing imputation, there were limited mechanisms in place to edit or check the consistency of the imputed data in the recipient record except for some very basic checks to verify that parts added to a required total. Problems with imputed data were mostly discovered by analysts using in-house developed analysis tools called Data Review (DR) and the Analysis System. This was due in part to the way the edit code was written, as it required DLT processing to finish before any imputation calls were resolved - in effect, the imputed data was not available to the module from which those imputed values had been requested. This inability, which didn't exist in 1997, has been resolved, so that in 2007 the imputed data will be edited by the same module which requested it, so that reported and imputed data will be edited by the same code.

39. For 2007, donor pool creation will begin with 2002 Census of Agriculture records. These records will be stratified using state, type of farm, size of farm and total value of production (TVP) into agricultural profiles that dictate a farm similarity between the recipient and all donors in the system.

40. Again, the Agency will make use of matching variables in the imputation process. But, this time, they will be used in a much more judicious manner. After a profile is chosen, a donor or donor data must be selected. This will be done using matching variables. In 2002, the list of matching variables for an item could have unlimited length. In 2007, the list of matching variables will be much shorter and will be created at the questionnaire section level and not at the questionnaire item level.

41. Imputation will take on a very new look for the 2007 Census of Agriculture. The research has been split into two very separate, but related pieces. The first piece is the creation of a pool of donors to find data for a recipient record. For a questionnaire the size of the agricultural census, very few records will go through the editing phase without needing some type of imputation. The question then becomes, how do we impute without a decent starting donor pool? The seeding of the donor pool will be with 2002 Census of Agriculture data. This data will be stratified using variables that separate records into profiles, where profiles define a high-level of similarity of farms. The Agency decided that physical distance should not play as important a role in finding a donor as it did in 2002. It is more important to impute data from records that are similar in type to the recipient record. These profiles will be defined using information such as the type of farm, the amount of land in the operation, and Total Value of Production (TVP) within the state of the recipient record.

42. Part two of the change in imputation methodology involves a change in the process that actually selects a donor from the donor pools. More emphasis is being placed on selecting a donor with data that preserves the already known distribution of the consistent census data. This would guarantee that the same record or group of records would not be used to impute over and over again. The imputation sub-routine will have the flexibility to choose a donor in a variety of ways depending on factors such as : available donor data, item and module needing imputation. One option is to randomly select a donor from a specific profile. Another possibility is to use matching variables that are module specific to find the nearest neighbor in the profile. Additionally, donor data could be created from a group of donors in the profile. This would be something like a weighted average or composite donor.

43. Quality control measures are the most important addition needed for the imputation piece during the 2007 Census of Agriculture. The modification to the DLT code that will edit the imputed numbers returned by the imputation sub-routine will help analysts to be more efficient. Since they will see the record only if it

fails the edit, they will most likely only touch a record one time during the 2007 processing. As part of the donor selection processing, certain other quality control measures are being considered. One imputation statistic that will be used will monitor imputed values to ensure that we do not impute the same value too often. The imputation procedure will also try to choose donor records for imputation that contain a minimal amount of imputed data. One additional quality assurance measure being considered is how to monitor the available donor records and to choose a donor record or donor data that would preserve the underlying data distribution of the records already processed.

D. Data Analysis

44. With so many changes and so many problems encountered, the in-house developed data analysis and review tools played a very important part in the processing of 2002 census data. There were two highly successful additions made for the 2002 Census of Agriculture. Data Review (DR) was used at the questionnaire level for microanalysis. It could be used to review and/or update census data as a stand alone tool or accessed from any number of analysis tools by double-clicking on the record's ID or a data point on a chart or graph. The Analysis System, used for county and state level (macro analysis) review, was designed to give the analysts the ability to review aggregated 2002 data, see changes between 2002 data and 1997 data, and drill-down to individual record data using DR to investigate possible data inconsistencies.

45. For 2007, the Data Review and the Analysis System will continue to be a very important and versatile tool for reviewing data. After the 2002 Census of Agriculture was completed, NASS asked its many data analysts for input about how to improve these systems and how to make them better editing tools. Analysts were also asked to share which features were most often used and which features were never used. Based on these responses, both analysis tools will be streamlined and the remaining code will be optimized to improve speed without a loss of functionality.

III. CONCLUSIONS

46. The most important lesson NASS learned from its recent experience with the Census of Agriculture is the importance of testing an entire system, and testing thoroughly, before actual processing begins. This lack of testing was a by-product of a series of poor choices. After the 1997 census, NASS decided to make wholesale changes in how the census is processed. By the time the publication was produced in 1999, it was already too late to begin the process ultimately used for the 2002 Census of Agriculture. With limited experience in organizing, implementing and administering such a census, NASS should have planned a series of changes to take place over a few censuses. Instead, the Agency made major changes in the areas of data capture, editing, imputation, and data analysis. The Census form itself also underwent wholesale redesign, with many new questions added.

47. Additionally, after the 1997 Census of Agriculture was published, the Agency failed to immediately begin planning for the 2002 Census of Agriculture. This poor planning was partly due to not having a decision-making mechanism in place to approve or disapprove of the planned changes. Delays at every step of the process planning plus wholesale changes that prohibited the Agency from being able to use 1997 data for testing, made meaningful testing for 2002 impossible. The entire 2002 experience consisted of deciding, writing, implementing just one step ahead of production processing - or sometimes concurrently with it.

48. To make sure this doesn't happen again, NASS has already taken many steps to facilitate a smooth and efficiently run 2007 Census of Agriculture. Agency-level managers have divided the 2007 census project into ten distinct areas and assigned a program manager for each one. These program managers meet every week to discuss processing timelines and resources needed for the 2007 Census of Agriculture. Everyone in the Agency is aware of the trouble and problems encountered during 2002 processing. All employees are dedicated to planning better and planning earlier; starting to work earlier; and to having all of 2007 for testing of the entire system. Inside NASS, we are famous for our 'we will get it done, somehow' approach, but better planning and design will help us to work on the 2007 census more efficiently and without the constant feeling of being behind schedule.

References

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

Atkinson, D. (2003), The Development and Implementation of a New Processing System for the 2002 Census of Agriculture, *UNECE Work Session on Statistical Data Editing*, Madrid, October 20-22, 2003.
