WP.40
ENGLISH
ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Vienna, Austria, 21-23 April 2008)

Topic (vi): Censuses

# FURTHER IMPROVEMENTS TO AN EDIT AND IMPUTATION SYSTEM FOR THE 2007 UNITED STATES CENSUS OF AGRICULTURE

**Supporting Paper**

Submitted by the National Agricultural Statistics Service, United States*

## I. Introduction

1. The National Agricultural Statistics Service of the United States Department of Agriculture (USDA-NASS) assumed responsibility for the quinquennial United States Census of Agriculture from the Bureau of the Census (BOC) in 1997. As has been recounted elsewhere [1][2], NASS implemented widespread changes to the entire editing process in 2002. These changes included converting the entire system from Fortran to SAS; introducing scanning technology and optical character recognition (OCR) of questionnaires for both data capture and storage/retrieval of the questionnaire image; significantly changing the questionnaire from the one used in 1997; implementing a donor pool and 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 and submit records for reediting. All these innovations were desirable and laudable; however, the resources, planning and testing required for the magnitude and number of the proposed changes were severely underestimated. Such overreaching meant much more time spent in the development of systems than had been anticipated, so that testing time was inevitably shortchanged. The result was a problem-plagued production process which necessitated numerous edit restarts and hundreds of thousands of reedits of records, with significant resource and data quality implications.

* Prepared by Dale Atkinson, dale_atkinson@nass.usda.gov and Jeffrey M. Beranek, jeff_beranek@nass.usda.gov

Some aspects of the changes made from 1997 to 2002 were very successful, but the overall success of the 2002 census processing was mitigated by the changes that were implemented less successfully.

2. Determined to avoid repeating these mistakes in the next census, NASS began planning for the 2007 census while the processing for the 2002 census was ending. In reviewing the deficiencies in the 2002 processes, NASS' upper management charged the Agency's staff with refining the processes for the 2007 Census of Agriculture to make them substantially better and more efficient. Among others, the areas needing attention included the database model, processing speed, donor pool and donor imputation methodology, and the general flow of the edit/imputation processing. Finally, it was imperative that all changes be implemented in time to allow adequate testing of the system and methodology before the mid-January 2008 start of data editing.

## II. Overview of the Editing and Imputation Process

3. In December 2007, NASS mailed census forms to just over 3 million operators on the census mail list (CML). An additional sample of 29,000 farm operators that were identified in the Agency's area frame and determined to be not on the mail list (NML) will be personally enumerated to provide a measure of list incompleteness. As in 2002, some list records will be reserved for computer-assisted telephone interviewing (CATI) or personal enumeration. In addition, NASS has implemented a Web-based electronic data reporting (EDR) option, which as of early February 2008 had accounted for approximately 40,000 reports. Returns are processed primarily at the Bureau of the Census' National Processing Center in Jeffersonville, Indiana, which prepares forms for image scanning and data entry, identifies and in some cases resolves problems, and transmits images and data to NASS. Data are formatted for editing and the records are run through edit and imputation in batches of 1,000 records. The goal is to process about 22,000 records per day.

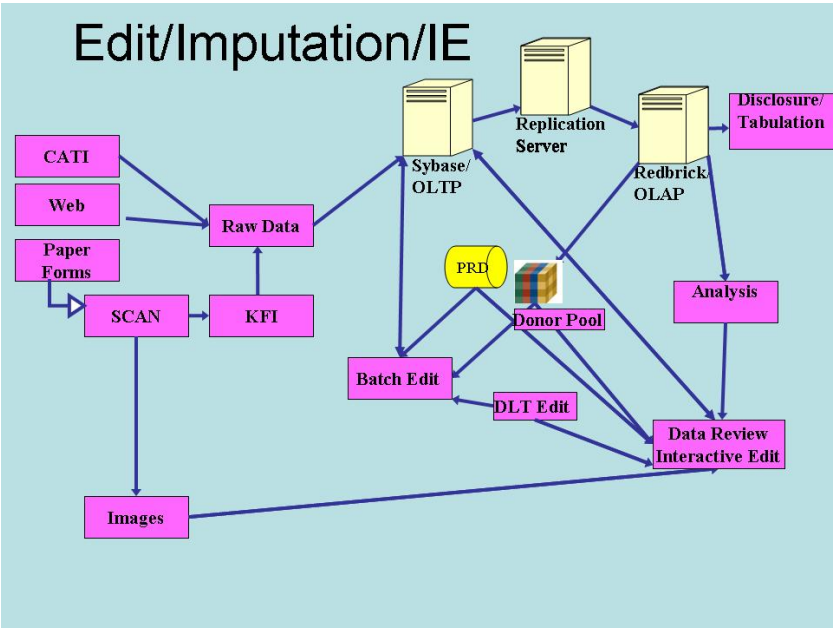
4. The edit sub-system itself consists of Decision Logic Tables (DLTs), which are 'if-then-else' logic expressions custom-written to check the census data. The DLTs are organized by module, where a module roughly corresponds to a section of the questionnaire. Missing data, or those judged inconsistent or incompatible with other data, will be altered either by deterministic logic (e.g., replacing a reported sum by the sum of the reported parts), by replacement with a (possibly adjusted) value from a previously reported survey, or by donor imputation from a similar clean record. Questionable relationships in the data will be noted using three indicators of increasing severity, and the records will be either reedited via the data review interactive edit, or marked clean and posted to the database.

## III. Improvements to the Processes for the 2007 Census of Agriculture

5. **Database Model**. The 2002 editing process required transactions for a single record (or batch of records) to be posted to two different databases on two different

servers -- Sybase, an online transaction processing (OLTP) database, and Redbrick, an online analysis processing (OLAP) database. One set of transactions consisted of the data themselves, while the other one consisted of associated administrative and system data (e.g., whether the record was clean, any error codes which had been set, and so on). Trouble ensued if, as too often happened, one set of transactions was posted while the other was not. In this case, an unedited and likely inconsistent record could be erroneously passed on through the system as a clean one. Moreover, since Redbrick is not specifically designed for transactional processing, every update to data ended up as a "load" to the database, often resulting in backlogs of hundreds of database loads competing against each other for an opportunity to post. Many of these were "deferred batches of one" of updates submitted from the data review screens in "interactive" mode. This database posting competition resulted in some interactive updates literally taking days to process.

6. The 2007 redesign still uses separate Sybase and Redbrick databases, but the roles of the two have been optimally redefined. All data editing and updates (transactional activity) will be posted to Sybase and all analysis will be done with Redbrick. Administrative data and reports will continue to be maintained in Sybase. A replication server will ensure that the data are synchronized between the two databases. By dividing the editing and analysis across the databases, the new design will take advantage of the strengths of each, competition for resources will be dramatically reduced, and analysis jobs will not interfere with editing jobs. The new design provides many advantages for improving the edit, including 1) access to all current and historical data, 2) a consistent data definition and metadata structure, 3) provision of an audit trail, 4) seamless connectivity among all modules in the system, including edit and analysis, 5) provision for an edit monitoring system at the micro and macro level 6) reduced duplication of effort and errors introduced into the system, and 7) scalability to accommodate growth in the number of users and quantity of data. The actual system flow is captured in the following schematic diagram [3]:



7. **Content and Form**. Considerably fewer changes were made for the 2007 questionnaire than were made for the 2002 form. The major change for 2007 was the addition of a 'Practices' section and the expansion of the existing 'Organic Agriculture' section. As was the case in 2002, a short form was designed to collect data from a sample of the records; however, the methodology differed in 2007. In 2002, the long form included sections of data that were not collected at all on the short form, and sample weighting was used to adjust aggregates for those "long form only" items. In 2007, no sample weighting will be performed since virtually all items are collected on both short and long form in a more or less detailed fashion. Specifically, the short form contains many open form questions (i.e., lacking the preprinted item codes that appear on the long form), and certain sections (such as production contracts) that do not apply to all farm operators are not included. The motivation behind developing the short form was to reduce cost and perceived respondent burden for smaller, less complicated operations with less agricultural activity to report. The monetary savings derived from printing and mailing only one shorter form of 12 pages, instead of a 24-pages region-specific long form, while the putative reduction of response burden was attained in providing respondents a physically shorter form to complete.

8. The differences between the long and short form are handled at the record level, either deterministically by the edit, or via imputation, so that the resulting edited record appears the same as it would have had the data been collected on the long form. To minimize the impact of the differences between the two forms and the amount of "record repair" required to accommodate them, the selection process for the short form excluded 'complicated' operations (e.g., operations with large livestock inventories or sales, nurseries, operators who had previously reported production contracts, etc). Approximately 500,000 operations were ultimately sent this form, and the edit/imputation system had to accommodate these and ensure that the final edited data collected on the short form would be identical structurally to those collected on the long form.

9. **Testing.** To ensure that the problems with processing the 2002 census data didn't reoccur, the agency implemented an aggressive testing plan for the 2007 census. Individual processes were upgraded and enhanced, where needed, and tested in isolation during 2006. Full scale integrated system and methodological testing began in February 2007. Editing of the 2005 Census Content Test data began in May 2007, and these data were reedited in December 2007, to ensure that they were clean with respect to the most current version of the edit. Reediting of mapped raw data from the 2002 COA took place over the summer of 2007. Finally, during the fall of 2007 stress testing of the system was performed to simulate typical processing loads expected during production.

10. **Data Capture.** For the 2007 COA, NASS eschewed the error-plagued OCR technology used in 2002 in favor of a technology called Key from Image (KFI) with Optical Mark Recognition (OMR). After the paper questionnaire is initially scanned into the system, the KFI technology is able to identify marks made in predetermined areas of the page, and to direct the key entry operators to just those cells. Marks made in check boxes are captured without keying. This process is much faster than keying from paper, and much more accurate than OCR. A second key entry operator independently keys a

portion of the data, and the software identifies the differences for later review. Tests run during development were very promising, showing differences in less than one half of one percent of all keyed items. The adoption of KFI made it possible to reintroduce the identification of simple edit failures into the 2007 COA. Simple edit failures are changes made by the respondent to the form that may indicate problematic keyed data. In the 2007 census, simple edit failures will be reviewed after the initial edit of a record whenever the value of production is at least $10,000. Simple edit failures were not identified in 2002 (though they had been in previous censuses), since the OCR software did not support this capability.

11. **Processing Speed and Batch Size.** In testing and early production work, batches of 1,000 are running through edit/imputation processing in about 3 to 4 seconds per record (50-65 minutes per batch), a great improvement over the hour or so required to run a batch of 75 records in 2002. Records edited using the Data Review utility (the tool for single-record, interactive editing) are returning within 5 to 15 seconds, thus enabling the user to develop an understanding of the edit effects of submitted changes on the data. The immense benefit of this type of edit training was not possible in 2002, when the results of attempted error corrections were often returned from the edit in a matter of days instead of minutes. For the 2007 census, editing with the Data Review utility is truly an interactive process, rather than the 'batch of one' processing of 2002. Additionally, with the 2007 process, data are not posted to the database until the user indicates he or she is finished editing the record. Users are thus able to investigate different scenarios by repeatedly changing data and rerunning.

12. **Decision Logic Tables.** As with the 2002 Census, NASS is using modular DLTs to edit the 2007 Census of Agriculture forms. The main change to the 2007 DLTs (other than those required by content changes and those resulting from edit analyses of 2002 results or 2005 content data) dealt with improving the implementation of donor imputation, and the connection between the DLTs and the donor search and imputation programs.

13. One of these changes was to pass control for specifying the imputation parameters to the DLT. For the modules that require imputation, the DLTs identify all variables needing imputation with a series of "IMPUTE ()" statements. These are accumulated prior to the issuance of a 'Get Donor ()' call. The IMPUTE () call passes certain parameters to the donor search routine, including 1) a target variable, identifying which data item needs the donor data; 2) a switch that indicates whether the returned donor data can be zero or must be positive; 3) a ratio variable used to scale the donor data (or a place holding "0", indicating no such ratio is to be used); and 4) a list of matching variables to be used in identifying the nearest neighbor, or a place holding "0". A matching variable place holder of "0" indicates that the matching variables are defined at the module level; that is, all imputation calls within the module will be satisfied by a single donor, whose proximity will be determined by the matching variables specified for that module. A list of matching variables in an IMPUTE () action implies that a separate donor call (most likely resulting in the receipt of data from a separate donor) will be issued for the specified target variable, and the matching variables specified in the IMPUTE () statement will be used to identify the closest eligible record.

14. Finally, constraints may also be passed to the donor search routine from the "Get Donor ()" statement. An example of such a constraint would be a requirement that milk cows are not positive (i.e. the donor is not a dairy) when imputing for different types of cattle.

15. Another important change to imputation for 2007 is that the imputed data are immediately accessible to the editing system (i.e., they are available to be edited by the module which called for imputation). This was not possible in the 2002 system, so for 2007 the DLT code had to be updated to 'validate' imputed data after imputation. In cases where the imputed data are still identified by the edit as violating edit constraints, an 'imputation validation flag' is set and the data are (possibly) adjusted.

16. **Donor Pool and Donor Search.** An intrinsic problem with using a donor pool is the question of how to initialize it. One solution is to delay using donor imputation until a sufficient number of records from the current survey have been edited and cleaned. This was the approach used by NASS in 1997 and the BOC in previous years. In contrast, for the 2007 COA, NASS opted to **'**seed' the initial donor pools with data from the 2002 census and the 2005 Census Content Test. This initial seeding of the donor pools for the 2007 census using previous census data was an option that had not been available in preparing for the 2002 census, since the 1997 census data resided in an incompatible format on an inaccessible computing platform.

17. The seeding of donor pools for the 2007 census was a multi-stage process. First, clean, edited data from the 2002 census was 'mapped' to account for content differences between 2002 and 2007. Some specific mapping issues that occurred included certain types of vegetables or field crops which were specified in 2002 and subsequently subsumed under 'other vegetables' or 'other field crops' in 2007, and conversely, items which were subsumed under 'other' in 2002 only to be uniquely specified in 2007. Fortunately, though, the content changes between 2002 and 2007 were fairly minimal, and this initial step of mapping the 2002 data to resemble 2007 data was handled without too much difficulty. The mapped 2002 census data were not reedited, but were used to create temporary donor pools used in building the Census Data Repository (CDR), which would in turn spawn the initial donor pools for editing the 2007 COA responses.

18. These initial 'production' donor pools consisted of approximately 12,000 clean, edited records from the 2005 content test plus about 270,000 clean, edited records from 2002, mapped to resemble 2007 content and edited with the 2007 census edit. In cases where both content test and previous census data were available for the same operator, the content test data were used, since these were more recent. The donor pools are stratified based on 1) physical location, 2) total land in the operation, 3) type of farm (one of 16 categories determining the major function of the farm), and 4) the total value of production. The strata are module specific, and some modules do not use all four variables in their donor pool stratifications.

19. The downside of using 'old' data to initialize the donor pools is the fear that the 'old' data may be dissimilar to current data, either in level or distribution. This was a

concern, for example, in imputing crop production and yields. To alleviate these concerns and avoid any such problems from imputing with old records, during the first few weeks of processing, records requiring imputation for production of field crops or hay are being halted and 'set aside.' Once a sufficient number of 2007 records with hay and crop production are edited and the donor pools are refreshed with current year data, then the set aside records will be edited against the more representative donor pools. Indications from the first few weeks of processing are that approximately 15 – 20 percent of records with crops are being 'parked' in this way.

20. **Matching Variables**. As mentioned above, these can be either module-specific or item-specific in 2007, depending on whether donor imputation is done by module, where one donor provides all data required for any imputations in the module, or by item, where different donors might provide data if, for example, both corn and soybean production were required to be imputed. This is a hybrid of the 1997 census imputation, in which all imputation was performed at the item level, and 2002, where all imputation was at the module level. To prevent the influence of any one matching variable from being 'diluted,' no more than six, and usually fewer, matching variables are used for any given donor search.

21. **Nearest Neighbor Calculations.** Similar to the approach for the 2002 COA, the nearest neighbor calculations for the 2007 census use an m-dimensional Euclidean distance measure, based on the matching variables. To account for differences of scale associated with the matching variables, the m matching variables are normalized by their variances before a distance is calculated. This is the special case of a Mahalanobis distance measurement in which the covariances between matching variables are all assumed to be zero. The imputation system, as programmed, is capable of accommodating the more general case, but testing prior to the 2007 census didn't show any significant quality gains from the more complex formulation, so the agency decided to stick with simple and fast. Mathematically, the formulation for the distance $\rho$ between recipient point R (with m matching variables) and D (a candidate donor) can be expressed as follows:

$$\rho_E(R, D) = \sqrt{\sum_{k=1}^{m} a_k (x_k - y_k)^2}$$

where $a_k > 0$, for $1 \le k \le m$. In our application the scale factor $a_k$ is chosen to be the reciprocal of $\sigma^2_k$, where $\sigma_k$, the standard deviation of the population of values associated with the $k^{th}$ coordinate. This choice of scale factor results in a normalized Euclidean distance and is motivated by at least two considerations:

(i.) it adjusts for different units of measurements for each coordinate in such a way that the resulting metric is physically meaningful and

(ii.) it weights the contribution to the distance measure from a given matching variable inversely by that variable's inherent variability.

22. Two modifications to the basic distance function were implemented for the 2007 census. These were to incorporate 1) a geographic distance between the recipient and potential donor and 2) additive penalty constants to minimize the use of less desirable donors. The relative sizes of the penalty constants are based on imputation statistics maintained by the system to monitor the desirability of candidate donor data. These statistics track the age of a potential donor's data and how many times its data have been previously donated. The idea is that newer donor data and data that haven't been previously donated are preferable to older data and data that have been previously used. The basic distance formula, incorporating the normalized Euclidean distance, the geographic distance, and the two penalty constants is of the following form:

$$\rho(R, D) = \sqrt{\rho_E(R, D)^2 + c\rho_G(R, D)^2} + A + U$$

Where $\rho_E$ is the normalized Euclidean distance between the m-dimensional recipient point R and potential donor D; $\rho_G$ is the geographical distance between the recipient and potential donor; c is a scaling constant to account for differences in measurement unit, variability and relative importance; A is the penalty constant for the age of the potential donor's data and U is the penalty constant for frequency of its previous usage. As of the writing of this paper, the exact formulation of the weighting and penalty constants was still being refined.

23. **Quality Control Measures.** To improve upon the 2002 process and ensure that data quality and processing issues are identified and addressed as early as possible, the agency has made a concerted effort to institute various safeguards. Specifically, the agency formed a team to address quality control issues, to ensure that the processing of the 2007 COA data is optimally efficient and effective. To carry this out, the team developed more than 30 different data quality tests that will be run regularly during the processing of the 2007 COA, with the output being routinely reviewed by headquarters analysts. With this monitoring in place, issues that might cause a system problem or an inconsistency in the data can be identified, and the edit can be halted and the code or procedure modified before the problem becomes too widespread.

24. The resulting new and integrated Quality Assurance/Quality Control (QA/QC) utilities have been introduced to ensure appropriate functioning of the system, as well as high quality data. The primary purposes of these utilities are to detect 1) lapses in progress in editing collected data, 2) system problems that could impact the data, 3) deterioration in system performance, and 4) problems in system stability. Such issues will be continuously monitored through the use of a dashboard. Problems deemed critical, or requiring immediate intervention will be traffic-lighted in red, while borderline problems will be indicated in yellow.

25. Additional important features of the QA/QC system include the following:

- Tracking and diagnostic capabilities to enable monitoring the impact of editing and imputation on total survey quality;
- Analytics to ensure that edited and imputed data are trusted;
- "Real time" reports on clean records that appear unusual;
- Provision of appropriate role-based access to system reports;
- Time series charts to track week-to-week progress;
- User friendly graphical interface with traffic lighting, critical zones, and drill-down capability.

26. Additionally, a quality check edit (DLT-based) of 50 or so basic relationships will be run periodically against a sample of clean records to reveal problems in either the edit or the databases.

## IV. Topics for Further Research

27. The development and testing process of the edit and imputation system for the 2007 COA has been generally considered highly successful, as the deficiencies of the previous census systems were addressed, and there was generally adequate time to react to new problems which arose as development and testing proceeded. Even so, more research is indicated in certain areas, including the following:

- Speed of data transfer from Sybase to Redbrick. Testing showed that this process became very slow if edit batches involve more than 1,000 records. Removing this limit would allow larger batches of data to be edited at a time.
- Donor imputation and donor pool seeding/weighting of donor records. While the 2007 process is far superior to that of 2002, the process in setting it up was quite labor-intensive. In addition, the fact that certain records were 'parked' until current data were available to be used in imputation underscores a drawback of using 'old' data to initialize the donor pool. More research into relative weighting of donor records based on age, number of times used, amount of imputed data on the record, etc., as well as alternative ways of stratifying the donor pools, and selecting donors from within the strata should also be undertaken.
- Editing of Demographic and Economic Data. It's widely accepted that our editing of these data needs to be improved, since the processes and methodology developed for editing crop acreages and livestock inventories are considerably less satisfactory for editing sales, expenditure data, and demographic information. The demographic section of the census form, in particular, can almost be seen as a separate survey appended to the census form.
- Reconsider the use of short and long forms. Explore the possibility of one form, shorter than the current long form length of 24 pages, to be used for all respondents.
- Research further refinements to edit code to enable the selective editing of records.

28. By the time the Work Session on Statistical Data Editing convenes in April, NASS will have been editing 2007 COA data for three months. During that time we will have had ample opportunity to observe the speed, efficiency and reliability of our edit systems, and to begin to analyze and measure the consistency and quality of the edited data.

## V. Conclusion

29. All aspects of the development of the edit and imputation process for the 2007 census ran substantially ahead of the 2002 pace. And, as painful as the 2002 process was, there were very valuable lessons learned from having been through it. As all census processing years are, we expect 2008 to be a very exciting one. However, after all the processing system improvements we've been able to implement and the volume testing we've been able to conduct in preparation for the 2007 Census of Agriculture, we're fully expecting the processing of this census to be a much more positive experience than the one five years ago.

## References

1. Beranek, J. and McEwen, R. (2005), Improving an Edit and Imputation System for the United States Census of Agriculture, *UNECE Work Session on Statistical Data Editing*, Ottawa, Canada, May 16-18, 2005.

2. Atkinson, D., Beranek, J. and McEwen, R. (2006), Further Improvements to an Edit and Imputation System for the United States Census of Agriculture, *UNECE Work Session on Statistical Data Editing*, Bonn, Germany, September 25-27, 2006.

3. Jacob, T. and House, C. (2007), Generalized Census Processing System at the National Agricultural Statistics Service, International Conference on Establishment Surveys, Montreal, Canada, June 18-21, 2007.