

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

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Bonn, Germany, 25-27 September 2006)

Topic (v): New and emerging methods

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

Supporting Paper

Prepared by A. Dale Atkinson (dale_atkinson@nass.usda.gov), Jeffrey M. Beranek (jeff_beranek@nass.usda.gov) and Robert P. McEwen (bob_mcewen@nass.usda.gov), National Agricultural Statistics Service, United States

I. INTRODUCTION

1. When the United States Department of Agriculture – National Agricultural Statistics Service (USDA-NASS) assumed responsibility for the United States Census of Agriculture from the Bureau of the Census (BOC) in 1997, few changes were made to the way the Bureau of the Census had conducted it in 1992. This was a direct result of time constraints. Approximately 50 BOC employees transferred to NASS to assist with the transition, and these folks were very instrumental in making the transition smooth.

2. After successfully conducting the 1997 Census and publishing the results, NASS began to plan for the 2002 Census. Many NASS employees felt that the processes and methodologies long used by the BOC would benefit from an upgrade. Many changes had to be made, since the 1997 Census had been run on aging hardware and software owned by the Census Bureau. Others were voluntarily made, as NASS' upper management charged the employees of NASS to 'think big' in planning for the 2002 Census of Agriculture (COA).

3. As a result, the agency made widespread changes to the COA process. Some of the changes made were: scanning questionnaires for capture and retention, instead of key entry and physically storing the returned paper questionnaire; nearest neighbor imputation instead of the hot deck methodology; adjusting for non-response and undercoverage instead of just non-response; and publishing the results a year earlier from date of mail-out than had been previously done.

4. There were some successes with the changes made from 1997 to 2002, but there were also some areas of change that were significantly less successful. In reviewing the deficiencies in the 2002 processes, NASS' upper management again charged the employees of NASS to make a better and more efficient process for the 2007 Census of Agriculture (COA).

5. One area of the edit and imputation process that was identified as needing significant improvement was the overall speed of record processing. In 2002, it could take as long as several days for a reviewed and manually corrected record to be returned to the analyst. Such excessive processing times

resulted in large part from processing log-jams and database synchronization problems that too frequently resulted in the processing system having to be brought down. The speed requirement for similar processing during the 2007 COA is 5-15 seconds. With many of the problem areas that plagued us during the 2002 data processing having been addressed in the new database model, and many efficiencies having been implemented in software code and hardware utilization, early testing indicates that the speed requirement will be realized.

6. The agency is also planning to spend more time on methodological research for identifying the best donor. Due to a lack of research and testing time, resulting largely from starting the system development too late, a rather simplistic approach for creating donor pools and a simple Euclidean distance calculation for nearest neighbor determination was used for 2002 COA processing. With the overall census planning and preparations for 2007 ahead of those of the comparable period for the 2002 census by a least a year, both of these processes will likely be improved upon, both processing efficiency and data quality-wise, for the 2007 census.

7. Another problem encountered in 2002 was inconsistent data being loaded to the database, which was often the result of inconsistent donor data being imputed into a record. Since imputed data were allowed to flow into records without undergoing the rigorous editing checks to which reported data were subjected, significant data problems remained in "clean" data records. These records created problems well beyond their individual contributions to census totals, since they were also used as donors for subsequent records. This will not happen in 2007. Donor data, once imputed into a recipient record, will be edited and required to be consistent before the record is turned loose and written to the database.

8. The basic editing approach used in 1997 was changed to the use of Decision Logic Tables (DLT) for 2002. DLTs are systematic, linear expressions that exclusively use 'if-then' type of statements to test and check census data. For imputation the agency made the switch from a hot deck approach to a nearest neighbor methodology.

9. Many of the proposed changes to the processing system for the census were made with NASS' survey programs in mind. The long-range plan was to transition the survey program processes and methodologies into the new systems and procedures being created for the census. The system changes included better hardware and software as well as introducing a relational database approach into the process. Since the magnitude of the census dwarfs all of NASS' survey programs, it was thought that if it worked for the census, then it would work for our surveys. Our experience in developing the 2002 system, however, made it very clear that the census isn't just a big survey.

II. EDITING

10. As with the 2002 Census, the USDA-NASS will use module specific Decision Logic Tables (DLT) to edit the 2007 Census of Agriculture forms. Basically, a module corresponds to a section of the questionnaire and DLTs are systematic, linear expressions that exclusively use 'if-then' statements to test and check census data.

11. The DLT approach to editing the 2002 Census of Agriculture forms proved to be one of the highlights of the changes made to the entire census process. From the many years of experience that the USDA-NASS has editing questionnaires from its sample surveys, many of the needed edits were already available. All that had to be done was to translate that logic into the DLT format. A team was formed to create these DLTs for the 2002 COA. Another one has subsequently been formed for the 2007 COA.

12. Since there were only minor content changes from the 2002 questionnaires to those for 2007, the 2007 DLT Team has had the good fortune of having a good starting point for their work. But, due to some additional procedural changes, this current version of the DLT Team has also had some new challenges to face in preparing for the upcoming COA.

13. One of these procedural changes is a new and improved call to the imputation routine. For the modules that require imputation, the DLTs will issue a 'Get Donor' call. In this call, it will pass certain parameters to the donor search routine. Among them is a target variable, identifying which data item needs the donor data; a switch that indicates whether the returned donor data can be zero or must be positive; and a ratio variable that would be used to scale the donor data to make it comparable to the recipient record being edited. The last parameter that may be passed to the donor search routine is a donor constraint. This would impose additional conditions on the donor search routine as it searches for the best donor data, such as requiring that a target or ratio variable to be greater than (or less than or equal to) a constant value or the value of another variable on the recipient or donor record.

14. One of the most important changes to the edit and imputation for the 2007 COA is that the recipient record will be re-edited after the donor data are returned to and inserted into it. In 2002 this was not the case, and inconsistent data were too often inserted into records through the donor imputation process, that would only be discovered much later in the processing.

15. If donor data are written to a recipient record and the edits still find problems with the record that cannot be automatically fixed, then the DLT issues a HALT statement and the edit stops. The record is marked to be reviewed by an analyst. An analyst will then review the record, make changes as needed, and resend it to the edit and imputation system for further processing.

16. During the analysis portion of the 2002 COA processing, many inconsistencies were found to be a direct result of imputed data that was not completely consistent with the reported data in the recipient record. Ensuring consistent donor data, by requiring them to pass the edits, will save many hours of data analysis and greatly reduce the need for the subsequent re-editing of records.

III. DONOR POOL CREATION

17. One of the big problems faced in 2002 was how to seed the donor pools for imputation. Since the questionnaire had been completely redone for the 2002 Census, and the 1997 data had furthermore been stored in a completely incompatible computer system, the 1997 data set could not be used to seed the initial donor pool. Fortunately, this is not the case for the 2007 Census. The current plan is to use all completed and consistent 2002 records as the starting donor pool for processing 2007 Census forms. There were some changes in the questionnaire from 2002 to 2007, but compared to the changes from the 1997 to the 2002 questionnaire, they are minimal. Those that do exist will be addressed in seeding the donor pools through the most appropriate methods available.

18. The agency prepared the 2007 Census questionnaire in 2005 and then proceeded to conduct a test of its content early in 2006. This content test was conducted nationwide with approximately 14,000 samples. It was designed to help the agency determine whether the questionnaire was well understood and the questions easily answered by farm operators, which is especially important, since the bulk of the census data are collected via mail. The agency was quite pleased with the results of this content test and made just a few additional minor adjustments to the questionnaire. Once these changes were made, the 2007 Census questionnaire was considered final. One other benefit of the content test was that it provided the agency with a data set of 14,000 additional records closely resembling the final 2007 questionnaire to use in starting a donor pool for 2007 census processing.

19. All records available to be donors will be stored in a Central Donor Repository (CDR). As the first 2007 records are being processed, the CDR will consist of the 2002 COA data and the 2006 Content Test data, all put in a consistent 2007 record format. Some atypical types of records, such as prison farms, university farms and agribusinesses, will be excluded from the CDR, since these would not provide data appropriate for imputation in most other records.

20. Any farm's data from the 2006 Content Test that are clean and consistent will replace (overlay) that farm's data from the 2002 COA in the donor pools, since the content test data are more recent and

they more closely resemble the 2007 COA data than do the 2002 COA data. The same line of reasoning will continue as 2007 COA records are processed. If a farm's data from the 2007 COA is clean and consistent, it will replace that farm's 2002 COA or 2006 Content Test data in the CDR.

21. The CDR will consist of a minimum of 1.5 million records at any one time, and it would therefore be very inefficient to search and perform calculations on all those records to determine the best donor. To provide the highest quality imputation results, while designing a system that is computationally efficient, the concept of farm similarity has been introduced into slicing the CDR into more manageable donor pools for the 2007 Census. The best donor will come from a farm that is similar to the recipient record (the one being edited), so a major challenge is to define parameters that will identify a similar farm. Since imputation will be done on a module-by-module basis, farm similarity must also be defined module by module. Therefore, for each module, a set of parameters have to be defined to identify farm similarity. The research to specify farm similarity specifications has now been completed in preparation for the 2007 census, and the master list of these parameters consists of state, total value of production (TVP), physical size of the farm and farm type. Farm type is the general description of the major function of the farm.

22. Using a module-specific combination of the farm similarity parameters outlined above, the CDR is partitioned into donor pools of a manageable size. The ultimate target size of these donor pools will be determined through additional research and is as yet undecided, but will most likely be around 100-200 records

23. When the DLTs determine that a record needs donor data, the search routine will therefore look for a donor solely in a pre-determined donor pool based on the farm similarity parameters. Pre-partitioning the CDR into these smaller donor pools will greatly reduce the computing workload.

24. In the donor search routine there is a built-in "positivity" requirement that ensures that any required ratio variable must be positive so that the donor data can be scaled to be comparable to the data in the recipient record.

25. Generally, the initial preparation of the 2007 production donor pool will be accomplished through the following steps: 1) Reformat the 2002 records to look like 2007 data; 2) Load these records into the CDR and create donor pools; 3) Map the 2006 Content Test data to the 2007 format and run those records through the edit and imputation code using the reformatted 2002 records (from step 1) as donor data; 4) Cleanup any "dirty" reformatted content test records and load the clean reformatted content test data into the CDR, overlaying any corresponding 2002 census data for matching farms; 5) Create donor pools for testing from the resulting initial production CDR.

IV. DONOR POOL MAINTENANCE

26. Every time a 2007 COA record is processed through the edit and imputation system, it will become an eligible donor for future 2007 COA records being edited. It is planned that a pre-specified number of records will be processed through the edit and imputation system every night during the data collection period. Once all of those records are edited and the edit shuts down for the night, the CDR will be refreshed. The newly edited 2007 COA records will be added to the CDR and will replace any 2002 COA records or 2006 Content Test data for the same farm.

27. As a specific way to monitor the records in the CDR, the agency will keep track of a set of imputation statistics. These will be used to assist in the final donor selection. There are three imputation statistics currently planned for use. The first is the year of the record in the CDR. It is more desirable to use a 2007 record, and these will be used whenever available. When that is not an option, then the 2006 Content Test data will be used. As a last resort, the donor search code will use a 2002 COA record as a donor. The second imputation statistic is an indication of the percentage of cells in a record in the CDR representing imputed data, since records with larger percentages of reported data are preferable as donors

to records which are largely imputed. The final imputation statistic currently tracked is the number of times a record in the CDR has already been used as a donor. It is undesirable to keep using the same donor over and over again for imputation. One important objective in using donor imputation is to preserve the underlying data distribution of the imputed data, and hence, the records in the CDR. By limiting the number of times a record can be used as a donor the underlying data distribution will be preserved.

V. DONOR SEARCH APPROACH

28. The creation of donor pools using the farm similarity parameters (outlined above) is the first step in finding a donor. The donor search program uses these farm similarity parameters to narrow its search within specially marked donor pools in the CDR. This is the Requestor part of the donor search routine. After the donor search routine finds a donor or creates donor data by compositing data from a group of donors from within the appropriate donor pool, it delivers this information back to the DLT via the Response part of the donor search routine.

29. In initiating the imputation process, the donor search routine accepts information from the Get Donor call issued by the DLT and accesses the CDR to begin its search for donor data. Once the donor search routine identifies the proper donor pool for the requested imputation, it begins calculating distances and statistics for each record in the identified pool and decides which donor data are the best for the recipient record. After the selection of donor data is made, the donor search routine accepts the information from the Requestor portion of the code and prepares the hand-off of the donor data to the DLT using the Response portion of the code.

30. The donor data are then written to the recipient record, and process control is returned to the DLT that issued the Get Donor call.

VI. MATCHING VARIABLES

31. Once the CDR is segmented into donor pools based on the farm similarity parameters and the appropriate donor pool is identified for a recipient requiring data, as requested by a donor call issued by a DLT, the best possible donor data for the recipient record must be found from within that pool. At this point in the process, matching variables are used to identify this “best” donor data. A set of matching variables is pre-designated for every item on the census that would potentially require imputation. This set consists of the group of variables in a record that is the most highly correlated with the variable needing imputation. The set of matching variables for a particular item will be used to calculate a distance from the recipient record to each record in the farm similarity donor pool. This is stage two for finding a donor.

32. Then, based on a pre-determined distance calculation, there will be one record in the farm similarity donor pool that will be the ‘shortest’ distance away from the recipient record. That record will be the nearest neighbor and supply the donor data required by the recipient record.

VII. NEAREST NEIGHBOR CALCULATIONS

33. For the 2002 COA, the USDA-NASS employed a simple n-dimensional Euclidean distance measure, based on the matching variables, to find a nearest neighbor for imputation. To account for differences of scale associated with the question specific matching variables, the n matching variables were normalized by their variance before a distance was calculated. This simple distance calculation, in addition to other possibilities, is also being considered for the upcoming census. Specifically, the agency is also exploring the possibilities of a Mahalanobis distance measurement and/or imputing donor data that are formed by compositing data from a small group of very similar nearest neighbors.

VIII. QUALITY CONTROL MEASURES

34. To improve upon the 2002 process and ensure that we can identify and address any data quality issues resulting from either reported data or data processing issues as early as possible, the agency is making a concerted effort to institute data quality safeguards. Specifically, the agency has formed a team to address quality control issues and ensure that the 2007 COA data set is of the highest possible quality. To date, this team has developed more than 30 different data quality tests that will run during the processing of the 2007 COA. These tests will be run on an on-going basis during the census and the output will be reviewed by headquarter analysts. If some procedure or process can be identified as the cause of a system problem or an inconsistency in the data distribution, then the edit can be halted and the code or procedure can be modified before the problem becomes too widespread.

IX. TESTING

35. The agency encountered many difficulties in processing the 2002 Census, most of which resulted from planning and testing that was inadequate for the huge revamping of the processing system that was undertaken. Testing was especially crucial for the 2002 Census since so many parts of the process were brand new, and so many others had undergone major changes. However, due to early under-estimates of the time and resources needed to revamp the systems, development time extended through the time that should have been used for testing the system.

36. System testing at even a small level would have revealed major problems with the speed of the edit and imputation (E&I) process. The 2002 COA effort suffered due to numerous systems related problems, but one of the biggest time consuming processes was the slowness of the edit in both the batch and interactive mode. This, coupled with the fact that the many records had to be reviewed by an analyst because the imputed data from the donor imputation process were not checked for consistency, resulted in a very onerous E&I process.

37. To ensure that the problems with processing the 2002 Census data don't re-occur, the agency has implemented an aggressive testing plan for the 2007 Census. Individual processes are to be upgraded and enhanced where needed and tested in a stand alone fashion during 2006. Already, in May 2006, the agency has executed a three module edit and imputation system test. This was designed to give IT developers an early look at how well the data and file hand-offs were being handled and to provide a check of the memory load on the system during production.

38. A full-scale system and methodological test of the edit and imputation system is planned for October 2006. This will provide the agency a good indication of the status and quality of the edit and imputation system a full 15 months before the system would need to be in place.

39. That leaves the entire year of 2007 for integrated testing of the entire processing system. The general plan is to stress the system and try to make it fail, by imposing unrealistically high volume testing. If the system can handle this type of volume, then it will be able to handle a normal amount of daily processing.

X. PROCESSING

40. All 2007 COA records will be edited as part of a batch editing process. Each night during the data collection period, about 20,000 records will be run through the edit and imputation system. The records will be edited using the DLT approach, and the donor search routine will be executed when needed. If the record passes all the edits or imputation can fix and/or complete a record, then that record

will be called consistent and will be written to the database. It will be available as a donor when the next batch of records is run through the edit.

41. If the record does not pass all the edits or imputed data are not consistent with the rest of the data in the record, then the record will be marked for review by an analyst, who will interact with the system through the interactive editing phase of the data processing. The analyst will review the record module by module, make decisions on the record and make holistic repairs to it. Once the analyst finishes editing the record, it must be run through the edit and imputation system again, essentially as a batch edit of one record.

42. Once every record is edited and consistent, then the edit and imputation phase of processing the 2007 COA will be finished.

XI. CONCLUSION

43. To date, every part of the development of the edit and imputation process for the 2007 Census is ahead of the 2002 pace. USDA-NASS has already corrected many of the issues with the 2002 Census processing system and appears to have dramatically improved the system, based on the limited testing done thus far. Early tests show that most processes are much faster and the hand-offs between different parts and different processes are working with fewer glitches. The agency is excited about what it can learn and trouble shoot from the aggressive testing that is currently planned prior to production.

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.

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