THE 2002 U.S. CENSUS OF AGRICULTURE DATA PROCESSING SYSTEM

Kara Perritt and Chadd Crouse
National Agricultural Statistics Service

Abstract

In 1997 responsibility for the census of agriculture was transferred from the Bureau of Census (BOC) to the U.S. Department of Agriculture's National Agricultural Statistics Service (NASS). Due to time constraints, the data processing system for the first NASS conducted census, the 1997 Census of Agriculture, remained essentially identical to that used by the BOC. The 1997 system emphasized efficient processing and powerful analysis capabilities, but lacked some desirable features including: seamless connectivity between processing components, graphical data analysis tools and a state of the art edit and imputation system. With these features in mind a new system is under development to be used for the 2002 Census of Agriculture and future NASS sample surveys. This paper outlines plans for two primary components of this new data processing system: the data edit and imputation system and the data analysis system.

Keywords: Edit, Analysis, Integrated, Interactive, Imputation

1. Introduction

In 1997 the responsibility for the five-year census of agriculture was transferred from the U.S. Bureau of the Census (BOC) to the National Agricultural Statistics Service (NASS) in the U.S. Department of Agriculture. This transfer motivated NASS to review its current survey data processing procedures and to evaluate the traditional census of agriculture processing systems. Opportunities to improve both the survey and census programs through effective integration were identified (Yost et al. 2000).

However, due to time constraints for the first NASS conducted census, the 1997 Census of Agriculture, few immediate changes to traditional census procedures and system were made. In fact, much of the data collection, data capture and editing was contracted out to the BOC's National Processing Center (NPC) in Jeffersonville, Indiana, which had assumed these functions in prior censuses.

The aging census system, basically unmodified since 1982, was run on DEC VAX machines with the VMS operating system. Users accessed this system via Telnet connections on their PCs. Use of Telnet denied users the familiar point-and-click navigation methods, and instead forced users to navigate an often-frustrating combination of keystrokes for record and aggregate level data review. In addition, both record and aggregate level review were entirely text-based, and the user had little control over the format or content of the review process.

The census edit and imputation program was relatively inflexible in that it was "hard coded" as a set of decision logic tables (DLTs) in Fortran. It employed a form of hot deck imputation; however, the donor pool was limited only to records in the current, single-state, edit batch. As records from the batch were processed, the most recent, usable data value for each item, within some predefined stratum, was stored for future imputation use. Cold deck values were seeded in the system for each

item until a usable value was identified. The combination of stratification and within-batch imputation resulted in multiple recipients using the same donor or cold deck value.

The text-based, census analytical review system used a two-phase approach to identify and correct anomalies at the aggregated county level. First, analysts reviewed approximately 7.4 million county totals (2409 items per county with 3078 U.S. counties) by comparing those totals to the previous 1992 census county totals. When an anomaly in a county total was identified, a sophisticated query system allowed the user to identify and tag specific, individual records potentially in error for that particular item. The second phase allowed the analyst to interactively edit records tagged in the first phase. However, the effect on the county total(s) of changing an item on any individual record was not made immediately apparent to the analyst.

In short, many opportunities to integrate and improve survey and census data processing procedures existed. Particularly, the entirely new data processing and analysis systems currently under development will feature the following:

- Improved edit and imputation routines;
- Graphical data analysis tools;
- Seamless connectivity among all modules.

This data processing system, to be used on NASS surveys and censuses, is being programmed in SAS®. The system will incorporate several enhancements to move toward a parameter driven, generalized system functioning through a Microsoft Windows™-based, point-and-click user interface.

As with the 1997 census, data capture activities for the 2002 census will be contracted out to NPC. Captured data, along with scanned images of each questionnaire, will be transmitted to the National Information Technology Center (NITC) in Kansas City, Missouri. The data and images will then be stored in a database and made available for processing through the edit, imputation and analysis systems. The flowchart in Figure 1 gives an overview of data flow from list preparation through dissemination. This paper will provide details on the topics shown in the shaded boxes of the flowchart.

2. The Edit and Imputation System

Traditionally, the NASS survey program relied on a manual review of all reported data by subject matter experts as the primary means of data editing. However, manually reviewing more than two million census records in less than six months is simply not feasible. Thus, groups charged with developing an edit and imputation system for the 2002 census initially planned for a "state-of-the-art" system which would:

- employ the methodology proposed by Fellegi and Holt (1976).
- be parameter driven to allow for easy integration with the NASS survey program.
- automate the edit process and minimize the volume of manual data review.
- feature one set of code for both interactive and batch editing.

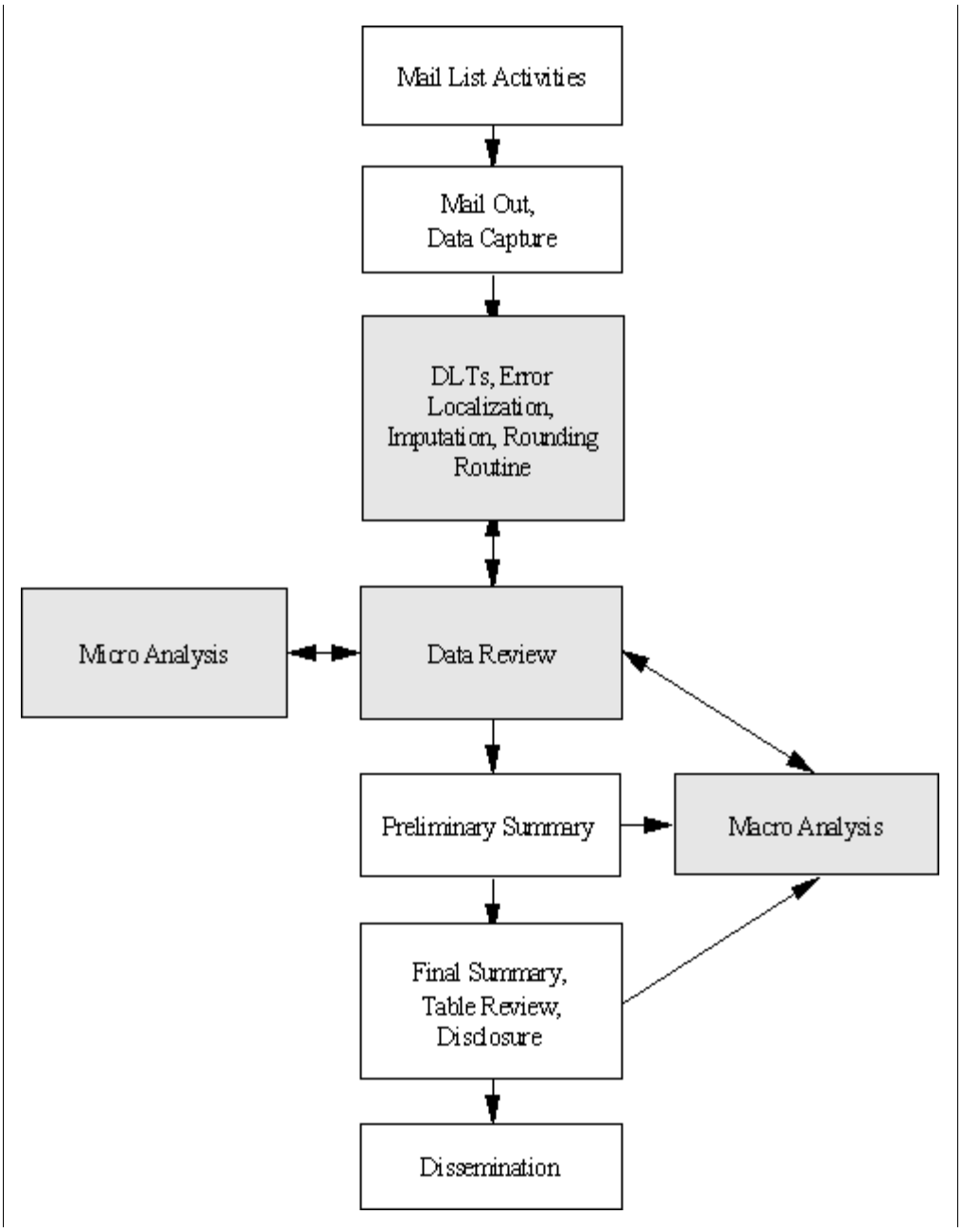

Figure 1: Overview of Data Flow

It soon became apparent that such a system had several drawbacks. First, the Chernikova algorithm (Chernikova, 1965), used in the implementation of Fellegi-Holt methodology, required a notoriously long processing time. Second, the automated minimal change philosophy, a principle of Fellegi-Holt, was not fully accepted in NASS culture where manual editing predominated. To overcome these drawbacks, a hybrid of the previous census system and a Fellegi-Holt-based system is being developed.

This hybrid system includes an automated edit and imputation program which is defined by a series of more than forty modules. Each module corresponds to a section of the report form and contains the following four steps which are also shown as shaded boxes on the flowchart in Figure 2:

- DLTs;
- Error localization;
- Imputation;
- Rounding routine.

Sections 2.1 through 2.4 of this report further define each of these steps. Before continuing however, some general comments should be made regarding the overall edit and imputation process. First, records are processed sequentially through all modules in batches. When data review (interactively editing one record) is performed, the batch-size equals one record. Second, imputation for any particular item is completed in one and only one module. Thus, item values established in a module cannot be changed in a future module. Finally, once a record passes through all modules, the record will be "clean, internally consistent and ready to publish."

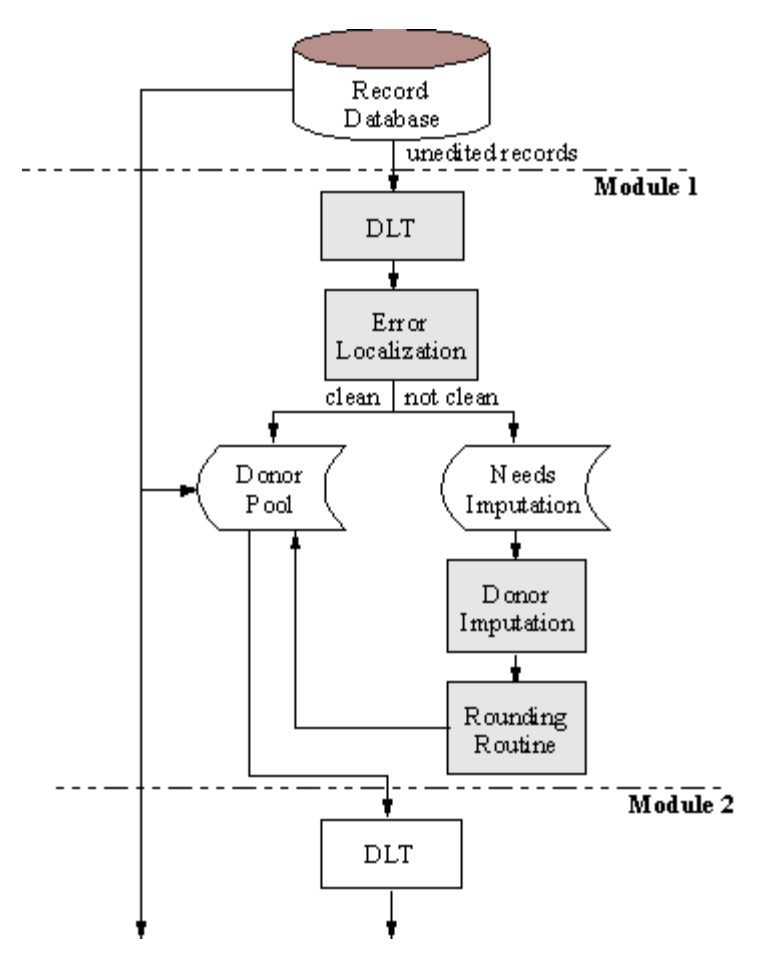

Figure 2: Flow of Data Through a Module

2.1 DLTs

DLTs consist of "hard coded" if-then-else logic statements which, when executed, identify and correct data inconsistencies. The new processing system under development uses DLT execution for the majority of data editing and imputation. This is not ideal since DLTs are neither parameter driven nor entirely objective; however, DLTs offer two substantial benefits. First, they allow subject matter experts some control over the imputation process, which is a compromise to the NASS practice of manual data review. Second, and more importantly, execution is very efficient; thus, overall data processing time is minimized.

DLTs are authored by subject matter experts who rely on the following three methods to correct data inconsistencies:

- Direct corrections;
- Corrections through the use of previously reported data;
- Corrections via donor imputation.

The first of these methods, direct corrections, is an attempt to correct problems based on the data available. For example, if the sum of beef cows, milk cows and other cattle does not equal the reported total cattle, the authors may decide to replace the total cattle value with the sum of the parts. This type of imputation is extremely efficient.

When the available data are not sufficient to correct problems, the authors may opt for the second method, the use of previously reported data (PRD). NASS contacts several thousand respondents each year as part of the ongoing survey program. Using these data in the edit and imputation process is intuitively desirable. For example, if a respondent reports only total cattle on a census form, the DLT authors may refer to PRD to impute the correct combination of beef cows, milk cows and other cattle. As mentioned previously, effective integration of survey and census programs was a principal goal in developing a new data processing system. The use of PRD in the edit process is the most obvious example of the benefits this integration provides.

When it is not possible to correct data inconsistencies through either direct corrections or the use of PRD, the DLT authors resort to the third method, marking specific items for donor imputation. When an item is marked for donor imputation, the authors must specify parameter values to indicate:

- whether zero is an acceptable imputed value.
- if the imputed value is to be pulled directly from the donor record, or
- if a ratio to some established variable value is to be employed (detailed in Section 2.3).
- the desired ratio variable if a ratio is requested.

Once an item is marked for donor imputation, no further editing on it may be done in the DLTs.

2.2 Error Localization

Immediately after exiting the DLT for a given module, all records pass through the error localization (EL) step. EL is an implementation of the Chernikova algorithm (Chernikova, 1965) which uses

a system of linear edits to check the internal consistency of each record. When a record does not satisfy all linear edits, the algorithm identifies, and marks for imputation, the minimal number of variables to change so the record can be made to satisfy all linear edits. The system of linear edits only checks items up through, and including, the current module. Weights are used to minimize the changes to values from previously edited sections. As EL marks items for imputation, the following imputation parameters are specified:

- Zero is an acceptable value;
- Data will be pulled directly from the donor record.

After records in a batch are processed through EL, they are categorized as one of the following:

- Clean - records passing all linear edits for the current module;
- Not clean - records requiring imputation.

Clean records are appended to the donor pool for consideration as a donor. Records requiring imputation are passed to the imputation step.

2.3 Imputation

All records requiring imputation first pass through a deterministic imputation routine. Deterministic imputation is defined as the case where only one value for an item to be imputed will result in satisfying all linear edits. The deterministic imputation routine initially executes for two reasons. First, locating a donor which satisfies all linear edits in these cases may be impossible. Second, the routine is significantly faster to process than donor imputation, which executes next.

The donor imputation routine first attempts to find the nearest neighbor:

- from ALL records (current or any previous batch) which are clean for the module.
- defined by the simple Euclidean distance of a predetermined, static set of variables.
- whose donated values will satisfy a set of relaxed linear edits.

The following are a few comments regarding nearest neighbors. First, using clean records from the current batch eliminates the need for seeded, cold deck values early in the process. Second, initial batches are grouped by state, and no state will be processed until a predetermined number of records have been collected for that state. This always ensures a sufficient pool of "reasonable" donors. Finally, latitude and longitude are always included in the predetermined, static set of variables. Thus, the nearest neighbor search is not limited to a single state or county.

After a nearest neighbor has been identified, the item values from that donor are applied to the recipient. As mentioned in Section 2.1, when specifying donor imputation in the DLTs, the authors indicated either direct imputation or ratio imputation. Direct imputation is straight forward; item values from the donor record are simply copied directly to the respective items on the recipient record. Ratio imputation is slightly more complex. In the DLT, a ratio variable is specified for one or more items to be imputed. The item(s) to be imputed from the donor record are divided by the ratio variable, and these ratios are multiplied by the ratio variable on the recipient record. The

results are imputed for the item(s) on the recipient record.

Whether direct or ratio imputation is used, the donor imputation routine checks the recipient record for internal consistency by applying a set of relaxed linear edits. If a record is not consistent, the next closest neighbor which will result in a consistent record is used.

Ratio imputation maximizes the likelihood of locating a suitable donor when the set of linear edits includes an equality edit, as in the cattle example where the sum of beef cows, milk cows and other cattle must equal total cattle. However, because most items are integer values, ratio imputation presents a minor problem when the rounded values do not exactly sum to totals. For this reason, the linear edits used in the donor imputation routine are relaxed, meaning the sum of the parts may differ from the total by as much as a rounding factor.

2.4 Rounding Routine

After donor imputation, records pass through a rounding routine as a final internal consistency check. In this routine, all item values are rounded to the specified precision (i.e., integers, tenths or hundredths), and the rounded data are passed through EL with the complete set of linear edits. Following EL, the records will pass through a deterministic imputation routine. Generally, a record which fails the set of linear edits does so because of an equality edit where the sum of the parts and the total are off by one. For these cases, the rounding routine randomly chooses one item from the equality edit and adjusts its value by one.

Records completing the rounding routine are internally consistent for every module up through and including the current module. Records then proceed to the next module. This process continues until the records in the batch have been processed through all modules.

3. The Analysis System

Records are available for review in the analysis system immediately after processing through the edit and imputation system. Note that all processes up to this point, with the exception of data capture, were completed without manual intervention. Thus, the analysis system will provide the first opportunity for analysts in the 45 State field offices to interact with the data.

The entire analysis system is being designed around interactive and dynamic tables and graphs. From any table or graph, analysts will select individual records to thoroughly review in an interactive tool, called data review. The data review tool is designed to allow for review and modification of any reported/imputed item values. Records modified in data review are reprocessed through the edit and imputation system, as a batch of size one, to ensure internal consistency. Once processed through the edit and imputation system, updated values will be reflected in refreshed tables and graphs in the analysis system.

To assure a sufficient data review, the analysis system supports the following two analysis phases:

- Micro-level analysis;
- Macro-level analysis.

Micro analysis, defined as any analysis of the data at the record level, begins immediately. Macro analysis - reviews of data at some aggregate level (usually county level for census data) - begins after a sufficient response rate has been achieved and preliminary weighting has been completed. Sections 3.1 and 3.2, respectively, discuss the micro and macro analysis systems in further detail.

3.1 Micro Analysis

The micro-level phase, which begins after the first batch of records has been processed through the edit and imputation system, involves the following three procedures:

- Correcting critical errors;
- Using graphs and listings to identify potential problem records;
- Reviewing high impact records as identified with a record score.

The first procedure, correcting critical errors, requires an analyst to modify some item value(s) in an individual record. Critical errors generally result when the edit and imputation system fails to completely process a record. This can happen for a variety of reasons such as: failure to find a suitable donor in donor imputation and/or the presence of some item value(s) on a record that exceeds a predetermined maximum allowable value.

By using graphs and listings, the second procedure, an analyst can easily identify specific records with potential problems, such as outliers or other data anomalies. The system will be preprogramed with specific graphs and listings intended to guide analysts through data review. Additionally, analysts may define ad hoc graphs and/or listings.

A score function is being developed to facilitate the third procedure, reviewing high impact records. The score, calculated for every record, will help identify records which represent a large percentage of the previous census' county total for preselected characteristics. These so called high scoring records potentially have the greatest impact on the county totals, and thus, may require additional attention.

3.2 Macro Analysis

After preliminary weights, adjusting for under-coverage and non-response, have been calculated, analysts will focus on the macro-level analysis phase. The system will calculate and maintain state and county totals by item in a tabular format. The main function of this phase is to ensure that every item in every county has been reviewed and approved. A check-off system will facilitate such a review. If a total does not meet approval standards, which are largely subjective, the analyst will drill down to the micro analysis tables and supporting graphics. From there, individual records may be selected for review and modification in the data review tool. Refreshed graphs and tables will reflect changes to the data.

Following the completion of data collection and the initial macro-level check-off, final weights will be calculated. One last review of the aggregates will take place prior to final summary and table review.

4. References

Chernikova, N.V., (1965), "Algorithm for Finding a General Formula for the Nonnegative Solution of a System of Linear Inequalities," *U.S.S.R. Computational Mathematics and Mathematical Physics*, No. 5, 228-233.

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

Yost, M., Atkinson, D., Miller, J., Parsons, J., Pense, R. and Swaim, N., (2000), "Developing a State of the Art Editing, Imputation and Analysis System for the 2002 Agricultural Census and Beyond," unpublished documentation, National Agricultural Statistics Service, USDA, Washington D.C.