STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 37
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27-29 May 2002, Helsinki, Finland)

Topic iv: Impact of new technologies on statistical data editing

DEVELOPMENTAL STATUS OF A NEW PROCESSING SYSTEM FOR AGRICULTURAL DATA

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

Contributed paper

I. INTRODUCTION

1. In 1997, the responsibility for the quinquennial census of agriculture was transferred from the U.S. Bureau of the Census to the National Agricultural Statistics Service (NASS) in the U.S. Department of Agriculture. This fulfilled a goal of NASS to become the national source of all essential statistics related to U.S. agriculture, and it provided an opportunity for the Agency to improve both the census and its ongoing survey and estimation program, through effective integration of the two. One near-term objective was to complete as much of the integration as possible by the 2002 census.

2. Following the 1997 census, the Agency took two major steps toward completing the integration. The first of these was the creation in late 1998 of the Project to Reengineer and Integrate Statistical Methods (PRISM). The team named to manage this project was charged with conducting a comprehensive review of all aspects of the NASS statistical program and recommending needed changes. The second step was a major structural reorganization of the Agency. This reorganization essentially absorbed the staff and functions of the Census Division, which had been formed specifically to manage the 1997 census, into an enhanced survey/census functional structure. The reorganization was designed to increase efficiency and to eliminate a duplication of effort by integrating the census responsibilities throughout NASS' organizational structure.

3. As part of this reengineering, there was a pressing need to develop a new processing system to standardize processing procedures for the census and NASS' sample surveys. There was an *immediate* need for this to be completed in time for use with the 2002 census. With the transfer of census responsibility, NASS had inherited an aging processing system that that was out-of-date technologically and, to some extent, methodologically. It sorely lacked the generality, user-friendliness and graphical capabilities that were deemed essential in a core survey data processing system – especially for use in NASS' decentralized processing environment.

---

[1] Prepared by Dale Atkinson (datkinson@nass.usda.gov).

4. As a result, in September 1999, the Processing Methodology Sub-Team of PRISM was chartered to specify a new processing system for the 2002 Census of Agriculture and subsequent, large NASS surveys. This group reviewed editing literature and processing systems used in NASS and other organizations (U.S. Bureau of the Census, 1996 and Weir, 1996) with the intention of synthesizing the best of what was already available into its recommendations for the new system. In February 2000 it published its findings and recommendations in an internal Agency research report (Processing Methodology Sub-Team, 2000). The report highlighted the team's guiding principles, as follows:

1) Automate as much as possible, minimizing required manual intervention.

2) Adopt a "less is more" philosophy to editing.

3) Identify real data and edit problems as early as possible

4) Design a system that works seamlessly with other NASS software and its databases.

5) Use the best features of existing products in developing the new system.

5. The sub-team documented the features it felt the new system should include, placing heavy emphasis on the visual display of data and on minimizing unnecessary review. It discussed display attributes and methodologies that could be used to identify problematic data with high potential impact on published estimates. The 'features' section of the paper discussed the need to refresh data analysis screens frequently as error corrections are made and for the system to help manage the review process (i.e., to identify previously edited records through color and/or special characters). The sub-team concluded its paper with the following recommendations:

i) To the extent possible, use Fellegi-Holt methodology in the new system.

ii) Have the computer automatically correct everything with imputation at the micro-level (i.e., eliminate the requirement for manual review).

iii) Utilize the NASS data warehouse as the primary repository of historical data and ensure that it is directly accessible by all modules of the new system.

iv) Design the system with tracking and diagnostic capabilities to enable the monitoring of the effect of editing and imputation. Develop analytics for a quality assurance program to ensure edited/imputed data are trusted.

v) Incorporate a score function to prioritize manual review.

vi) Provide universal access to data and program execution within the Agency.

vii) Ensure that the system is integrated into the Agency's overall information technology architecture.

viii) Make the system generalized enough, through modular design, to work over the entire scope of the Agency's survey and census programs.

ix) Enable users to enter and access comments anywhere in the system.

x) Present as much pertinent information as possible on each screen of the system and provide

on-screen help for system navigation.

xi) Consider the use of browser and Java programming technology to assist in integrating parts of the system across software, hardware, and functions.

xii) Designate a developmental team to take this report, develop detailed specifications and begin programming the system.

II. THE SYSTEM DEVELOPMENT

6. To begin the developmental work, a number of working groups were formed, each of which was to focus on one of the primary functional modules of the processing system. These included check-in, data capture, edit, imputation, weighting, analysis, and data review screens. In order to ensure consistency of decisions among the working groups, the Processing Sub-Team was formed as an oversight and technical decision-making body. This sub-team consisted primarily of the leaders of the individual working groups and was charged with ensuring the overall system flow, consistency, integrity and connectivity of the new system. The sub-team also served as the technical decision-making body for crosscutting decisions that couldn't be made by the individual working groups.

7. Through the tireless efforts of their leaders, the working groups gradually completed the often-frustrating specification writing phase of the system development. The early efforts were especially difficult for a number of reasons including 1) initial underestimation of the magnitude of the undertaking; 2) a very short time to accomplish the very ambitious undertaking; 3) the shortage of full-time staff assigned to it, with too many collateral duties of working group members; and 4) significant philosophical differences at all levels of the Agency in how to address important aspects of the system development – resulting in many early, time-wasting reversals of course.

8. In spite of the obstacles, however, the specifications writing process is now complete for most modules and NASS has entered the second (programming) phase of the PRISM project. On entering this phase, it implemented several important changes to its process management. Most of the working groups and the Processing Sub-Team itself were restructured to re-focus the efforts on programming the new system, and a PRISM Program Manager was named to oversee the overall PRISM process. NASS also began contracting with outside consultants to provide guidance in system development and to assist with the project management. The combination of these efforts and the immediacy of the 2002 census, which forced the Agency to rectify most of the situations that had led to the delays in Phase I, have resulted in encouraging progress in Phase II. Project momentum has increased significantly over the past year.

9. When finished the system will consist of a sophisticated interaction of data bases – Oracle (for image capture), Sybase (for transaction processing) and Redbrick (for historical data access) – with client/server application software written primarily in SAS. The following sections describe plans for selected modules of the system and discuss the progress made to date.

A. Data Capture

10. As was the case in 1997, NASS is contracting the printing, mailing, check-in of questionnaires and the data capture activities to the Census Bureau's National Processing Center.

However, while all data capture for the 1997 Census of Agriculture was accomplished through key-entry, the 2002 data capture will be primarily through scanning and optical/intelligent character recognition (OCR/ICR). Questionable returns will be reviewed, with erroneous data re-entered by correct-from-image key-entry operators. The only other key-entry of data, though, will be for operations requiring special handling by the Agency's State Statistical Offices (SSOs). Data for these operations will be entered through Blaise instruments. The scanning process will produce data and image files, which will be sent to a high-end, host Unix box for further processing. The data will pass into the editing system and the images will be displayed with the captured data in the data review screens.

B. Micro-Edit /Imputation

11. As the edit groups began to meet on a regular basis the magnitude of the task of developing a new editing module for the census became obvious. The machine edit/imputation used for the 1997 census was enormous! It had consisted of 54 sequentially run modules of approximately 50,000 lines of Fortran code, and the volume of the input decision logic tables (DLTs) was staggering. Through 1997 the census questionnaires had changed very little from one census to the next, so the DLTs and Fortran code had required little modification. For 2002, however, the new processing system had to support a questionnaire that was also undergoing radical changes – some due to recent structural changes in agricultural production and marketing and others due to the planned use of OCR/ICR for data capture. As a result, the group members were saddled with the onerous task of working through the mountains of DLTs from 1997 to determine which historical routines were needed in the new system, in addition to developing entirely new routines for some sections of the questionnaire.

13. Early plans for the 2002 census called for a statistical editing approach based on Fellegi-Holt methodology, with serious consideration given to the AGGIES developmental work in NASS' Research and Development Division (Todaro, 1999). However, the initial attempt to develop such a system was challenged by a series of organizational and technical obstacles which will be discussed in the following paragraphs.

14. The technical obstacles resulted from the length and complexity of the census questionnaire and the inefficiency of running traditional Fellegi-Holt based error localization on extremely large records. The Chernikova Algorithm is known to be very computationally intensive when applied to records with large number of variables and complex variable interactions. The census of agriculture data certainly qualify in this regard, coupling these problems with the potential of a very large number of records to edit. Indeed, the initial testing of the SAS-based error localization routine of AGGIES, using a mainframe computer and 1997 census data, was discouraging. And since initial plans were to perform most of the processing for the census on a mainframe at USDA's National Information Technology Center in Kansas City, the poor performance of the SAS routine on this platform was especially alarming. Even some of our older Unix-based workstations were running the error localization routine more quickly. The resulting efficiency concerns were exacerbated by delays in getting the overall system in place, so that the routine could be tested in context.

15. The organizational obstacles to the extensive use of a Fellegi-Holt based edit system arose as a result of the NASS editing culture. Conducting a census is new to NASS, and managing census processing is significantly different from managing the sample-based survey program to which it's accustomed. Sample sizes for its long-standing sample surveys are in the hundreds and thousands –

not millions, as they are for the census. As a result of its history of dealing with much smaller numbers of reports, NASS' culture has been one of touching every survey record. Each report has typically undergone both manual and machine editing – never being very far from a statistician's watchful eye. Even the machine edits have historically *only flagged* errors for review. A statistician would have to post an edit transaction to actually change a record's data.

16. While NASS staff generally recognize that this completely hands-on editing approach is not viable for the census, the thought of 80 percent or more of the questionnaires moving directly to the machine edit, being automatically corrected as necessary, and being passed along without manual review is uncomfortable to many in the Agency. Substantial discomfort with the processing of a large percentage of the 1997 census data without human interaction was expressed by Agency staff in post-census comments. Furthermore, insofar as the machine editing process for 2002 is viewed as a "black box," these concerns could be even worse for 2002. In particular, the heavy use of error-localization and donor imputation has the potential of raising additional red flags, in terms of further obscuring the source of machine-altered values. Coupling this with the prospect of even more automated changing of data in the upcoming census, with less manual review (due to an expected shortfall of staff for editing of the 2002 census relative to 1997), exacerbates the concerns.

17. The need to address the efficiency concerns and cultural resistance to a full implementation of statistical editing became a primary underpinning for the final edit and imputation plans for the 2002 census. The resulting implementation will make heavier use of direct IF-THEN-ELSE logic from Decision Logic Tables, and correspondingly lighter use of error localization, than initially planned. The use of donor imputation will also be reduced somewhat, as there will be substantial direct imputation performed in the DLTs. A review of the 1997 data in light of the plans for 2002 indicates that the vast majority of the imputation performed will be deterministic (e.g., forcing subparts to equal a total). Deterministic imputation could amount to 70-80% of all imputation for the 2002 census. Nearest neighbor donor imputation will likely account for 10-20%, while direct imputation of historical data (through the DLTs), perhaps 5-10%.

18. The heavier DLT usage addresses both of the above described concerns, since it is computationally more efficient and more comfortable to statisticians in NASS. However, an error localization/donor imputation presence was retained in the final plans, leaving the door open to a heavier use of statistical editing in subsequent surveys and censuses. The following paragraphs will describe the structure of the new editing system with a particular focus on its planned application to the 2002 census.

19. The micro-edit/imputation system for the 2002 census will consist of 43 modules, each of which addresses a different portion of the questionnaire. Each module consists of a DLT and module-specific error localization and donor imputation routines. For each module the DLT is processed first. A SAS macro then assesses each record coming out of the module's DLT to appropriately channel it into one of the three possible processing channels. The record will be processed as one that 1) must go through error localization, 2) may skip error localization but must go to donor imputation, or 3) may skip both error localization and donor imputation. Records in the third category are added to the donor pool for that portion of the questionnaire and are passed on to the next module.

20. Error localization will be cumulative. That is, all linear edits pertaining to the current and all previous modules will be included. Variables involved in earlier modules will be more heavily

weighted than those in the current module, to minimize the incidence of changing values that were earlier determined to be acceptable. This differential weighting improves the efficiency of the error localization process.

21. Donor imputation for a module will be performed by forming a donor pool of the best matching records, based on the match variables designated for that module. These best-matching records will be those with the lowest values of the distance function, which is computed as a sum of squares of distances between the normalized match variables. The pool of best matching records will then be ordered by distance and the recipient record will be imputed from the first candidate whose values pass all linear edits. Thus the nearest acceptable neighbor is used as the donor.

22. The modules are processed in a very specific order to ensure that all data from a previous module that are needed to edit items in the current module will be complete and "clean" prior to beginning the module. The entire micro-edit/imputation system is tied together with a "wrapper" program.

C. Macro-edit

23. The macro-editing portion of the new system is perhaps the module of interest to the broadest audience in NASS. This module, referred to as the analysis system, will provide the tools and functionality through which analysts in Headquarters and the SSOs will interact with the data. All processes prior to this point are ones with little or no manual intervention Because of the broad interest in the analysis system and the anticipated large number of users of it, the development team has made a special effort to solicit user input into its specification. The working group chartered to design and program this module circulated a hard-copy prototype of the proposed system to staff throughout the Agency early in 2001 and has since provided a series of demos to various Agency staff as enhancements are made to the system.

24. The macro-editing or analysis will consist of two basic phases – micro-analysis and macro-analysis. Each phase which will kick in at the appropriate point of processing the data. The phases are described chronologically in the following paragraphs.

25. After the data have been processed through the edit and imputation steps, during which essentially all critical errors have been computer corrected, they are ready for SSO analysis review. The first of the two analysis phases, micro-analysis, begins immediately. During micro-analysis SSOs will review (and update, if necessary) all records for which imputation was unsuccessful, any records failing consistency checks, and all those with specific items that were flagged for mandatory review. All records in one of these categories are said to contain critical errors and must be corrected. This work will be done while data collection is ongoing, and will allow ample time for any follow-up deemed necessary.

26. The system will also provide the capability to identify and review records that have no critical errors, but may be nonetheless of concern. Some "potential problem" records will be identified through micro-level graphical analysis. Micro-level graphics provide record level information distributionally for specific item(s) of interest, to show comparisons with other reports. The user has the option of subsetting the graphics by selecting a group of points or by specifying a subsetting condition. For some plots, the option of additional grouping and/or sub-grouping of a variable(s) through the use of colors and symbols will be available (e.g., by size of farm, type of

operation, race, total value of production, other size groups). Scatter plots, box-plots and frequency bar charts of various types will be provided, and all graphics will provide drill-down capability to data values. Data review screens will be used to review and update erroneous records.

27. Additionally during micro-analysis, the computer will identify influential or high scoring records. A score function has been developed for the 2002 census to ensure that records expected to have a substantial impact on aggregate totals are manually reviewed. Since county level aggregates are published, and are a product of special interest to the users, these are of particular concern in reviewing the census of agriculture data. Therefore, the score function used for 2002 will be county aggregate based. In particular, it assigns high scores to records whose current report represents a large percentage of the previous census' county total for that characteristic. The record's overall score is aggregated over a set of selected characteristics.

28. Finally, the system will track IDs that have been previously reviewed, compare current values to historic data, allow for canned and ad hoc queries and have a comments feature to document actions. Micro-analysis will also include tables to review previously reported data for non-responding units. This will allow SSOs to focus nonresponse follow-up efforts on the most "important" records. If micro-analysis is effective, the number of issues to be dealt with in macro phase will be greatly reduced.

29. The second phase of analysis, macro analysis, begins immediately after preliminary weighting (adjusting for long form sampling and non-response). Macro-analysis uses tables and graphs to review data totals and farm counts by item, county and state. The macro-analysis tools will retain the key objectives of the analytical review system used for previous censuses, but it will be much more interactive and user-friendly. The focal point of the macro-analysis will be a collection of graphics showing aggregate data at state and county levels. These graphics will include dot plots or bar charts of county rankings with historic comparisons, state maps with counties color-coded by various statistics and scatter plots of current vs. previous data.

30. The new macro-analysis tool will also be integrated more effectively with the Agency's data warehouse and its associated standard tools for ad-hoc queries. Graphics or tables will be used to compare current census weighted totals and farm counts against previous census values and other published estimates. Analysts will drill down to the data review screens to verify/update records

31. The macro-edit can be run as soon as the majority of data collection is complete, the records are run through edit and imputation, and preliminary weights are available. The objective of the macro review will be the same as for previous censuses. That is, the analyst will be responsible for the complete review of all the state and county totals. Every item in each county must be reviewed.

32. After all data have been "cleaned" from the preliminary phase of weighting and analysis review, final non-response and sample weights (to adjust for items only collected on the long form) will be generated. There will also be coverage adjustment weights generated at this point to account for incompleteness of the census mail list. All weights – sample, nonresponse and coverage – will be integerized and applied. The application of these final weights will spawn the final analysis review phase. At this point, a check-off system in the analysis module will be activated to ensure that all county totals that will be published are reviewed and approved by the SSO analysts.

III. CONCLUSIONS

33. As we enter the stretch run of putting the new system in place, much has been accomplished, but much remains to be done. The journey has been difficult, as we've navigated technical, cultural, timing and staff morale issues. At present the developmental progress is still considered to be behind, but we're adjusting as needed and making up some lost time

34. One concern that has been present throughout the developmental process has been over whether the ultimate performance of the system will be acceptable. Delays in getting to the point of testing the system as a whole have only intensified the performance concerns. However, the Agency is working to address the performance issue from both a software and hardware standpoint. As outside SAS consultants are working with the processing teams to optimize the performance of the various SAS modules as they are developed, NASS is purchasing the fastest available hardware on which to run the application. A 32-processor UNIX box with 128 *gigabytes* of memory will be used as the server platform for the system.

35. In conclusion, when our PRISM Program Manager was asked in early March 2002 how he felt about the status of system development he said "Well, things seem to be moving along right now. It would have been nice to be where we are today a year ago, but be that as it may, we'll get there from here." And we will, because that's the NASS way. We have a way of challenging ourselves by perhaps making things more difficult than they need to be, but we always seem to ultimately come together as an Agency to meet the challenge. That's the way we've always done things, and we expect the results of this undertaking to be no different!!

References

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

Fellegi, I.P. and Holt, D. (1976), A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, March 1976, Vol. 71, No. 353, pp. 17-35.

Processing Methodology Sub-Team (2000), Developing a State of the Art Editing, Imputation and Analysis System for the 2002 Agricultural Census and Beyond, NASS Staff Report, February 2000.

Todaro, T.A. (1999), Overview and Evaluation of the AGGIES Automated Edit and Imputation System, *Conference of European Statisticians*, June 2-4, 1999, Rome, Italy.

U.S. Census Bureau (1996), StEPS: Concepts and Overview, Technical report from the U.S. Census Bureau.

Weir, P. (1996), Graphical Editing Analysis Query System (GEAQS), *Data Editing Workshop and Exposition, Statistical Policy Working Paper 25*, pp. 126-136, Statistical Policy Office, Office of Management and Budget.