1982 ECONOMIC AND AGRICULTURE CENSUSES: AUTOMATION AND OTHER IMPROVEMENTS

Michael G. Farrell, W. Joel Richardson, John R. Wikoff, Bureau of the Census

To conduct the economic and agriculture censuses concurrently for the first time since they were last taken as part of the decennial census in 1940, fundamental changes to traditional processing methodology were introduced for the 1982 censuses. These changes are enabling the Census Bureau to process the combined censuses using approximately two-thirds the number of temporary employees previously required to process the two programs separately. With data collection and initial processing nearly complete, it is now clear that these changes will permit the release of the results of both censuses several months earlier than ever before.

Application of state-of-the-art computer technology, use of bar codes on labels and high-speed laser sorters, and use of microcomputers reduced or eliminated many of the labor intensive, repetitive tasks. Changes to the keying methodology dramatically increased document throughput, reduced keyer error rates, and eliminated the need for a separate prekeying screening operation; and innovations in questionnaire design reduced respondent burden, increased the rate of response, and dramatically decreased the volume of incoming correspondence.

BACKGROUND AND PURPOSE

Automation of the Census Bureau's economic and agriculture census operations has progressed continuously, but at an uneven pace, since the development of the punch card tabulating system by the Bureau's Herman Hollerith for the 1890 Decennial Census. Subsequent landmarks included the use of administrative records as control lists in the early postwar years to conduct the censuses by mail and the use of Univac I for the 1954 Economic Censuses [1]. Through the 1970's, subsequent advances consisted primarily of greater use of administrative records to construct proxy reports for small firms and the use of faster computers. There were few innovations in other areas. The 1977 Economic and 1978 Agriculture Censuses were processed much as they had been a generation earlier.

Data users increasingly were asking for earlier release of the data for benchmarking important economic indicators and for the many other purposes for which census statistics are used. Despite well-intentioned goals, there was little improvement during these years in release dates of census results. Users also were requesting a common reference year for the economic and agriculture censuses to link measures of input of one sector to activity in another. The law that provided for concurrent economic and agriculture censuses provided for moving up each agriculture census 1 year until the census years coincided for 1982. Concurrent processing could not be accomplished merely by extending the processing period. The goal for the 1982 Economic and Agriculture Censuses was established as concurrent processing with a 2- to 6-month reduction in the publication release schedule. This had to be accomplished without compromising quality.

SUMMARY OF ACTIONS

Planning for concurrent processing began in 1979. A thorough analysis was made of document handling techniques used in the 1977 Economic Censuses and the 1978 Census of Agriculture to determine where delays had occurred and what operations needed to be improved. Areas in which critical delays in processing occurred were identified as (1) competition with other Bureau programs for computer priorities, (2) batch sequential processing of administrative records, (3) the labor intensive coverage and search operations, (4) the check-in operation, and (5) screening of reports.

Other areas were identified that also offered significant opportunities for cost reductions and improvement of quality. But concurrent processing of the two programs along with improvements in timing required that the elimination of these major bottlenecks be given first priority. Examination of the problems of inadequate computer capacity, the limitations of batch sequential processing, the delay in processing administrative records, and the delay inherent in the coverage function (the functions required to maintain a current, complete, and unduplicated list of all establishments together with records of their owning or controlling firm) resulted in the conclusion that a dedicated computer to process documents and records on a flow basis using interactive applications programs for coverage and related records search functions was the best way to eliminate the major bottlenecks. Sufficient labor intensive operations were eliminated to finance the computer out of program funds. Additional areas for improvement or automation with a dedicated interactive system also were identified. Most importantly, functions were automated during the development of the resulting Census Control System (CCS), which otherwise would have resulted in new weak links in the processing system.

To provide greater flexibility in processing the censuses for Puerto Rico, a microcomputer network was developed and all pretabulation processing operations were performed on-site at the Bureau's temporary San Juan office. This action eliminated conflicts with computer processing of the main censuses, which had previously caused major delays in this auxiliary operation. Improvements in the planning operation made major contributions. A computerized time schedule assured that essential planning functions were completed in time. Improvements in forms design and instructions resulted in reductions in respondent burden and--equally important--of perceptions of respondent burden. In turn, this resulted in earlier response and reduced follow-up and correspondence costs. Incoming correspondence declined more than 60 percent from the previous censuses.

Among the other problem areas identified, those with the greatest potential to speed processing and reduce costs also were automated or sharply improved. Bar codes were used with the form mailing labels to automate the verification

of mail packages of multiestablishment firms. More importantly, this permitted automation of the check-in operation. High-speed sorters with laser scanning capability were used for initial sorting of the incoming forms and to identify schedules to receive special attention because of their impact on the published data. Improvements in methodology and training resulted in major improvements in the data entry, or "keying," operation; and screening of documents prior to data entry was eliminated through changes in keying procedures.

Two significant areas of improvement were in the publication area. Legal requirements to protect the confidentiality of data reported by individual firms require complex and time consuming procedures for identifying and suppressing potential disclosures in publication-level data to prevent the possibility of obtaining data for individual firms either directly or by comparing or subtracting data from different publication cells. The previous computer programs for this operation were improved to eliminate almost all manual intervention. A computerized photocomposition system to produce publication quality statistical tables had been introduced for standard table formats for the 1977 censuses. For 1982, this system was improved to provide for composition of all publication formats and to insert corrections, footnotes, and other necessary alterations to publication copy.

DETAILED ACTIONS TO IMPROVE AND AUTOMATE CENSUS PROCESSES

Interactive Computer Processing--The Census Control System (CCS)

1. Design of the system--In July 1981, a contract was awarded for the CCS computer, which fulfilled the following basic hardware and software requirements: 16 billion characters of on-line mass storage, 200 interactive computer terminals, 8 million characters of main memory, system security software, and automatic data base recovery capabilities.

Two years before the Bureau knew which contractor would win the award, work began on the design of the data base and on the access routines. Independent access strategies were developed well ahead of the award of the contract. A COBOL library containing hundreds of standard routines was developed to search, read and write on-line data records, perform validation edits on each data item, format data base update transactions, and perform system security and operator access validation checks.

Batch processing was speeded by the elimination of time consuming file sorting and segmenting by the direct access capabilities of the Data Base Management System (DBMS). Programming requirements common to several of the Bureau's subject divisions were combined and developed centrally to achieve further economies in programming and processing.

Over 250 interactive programs were developed to support the centralized processing activities of economic and agriculture censuses, the annual company organization survey and the annual survey of manufactures. Controlled efficient interactive processing was achieved using the following techniques: interactive processing is menu driven, all interactive procedures are program controlled, data are displayed in easy-to-read formats, screens can be scrolled forward and backward, companies or establishments that satisfy operator specified criteria can be retrieved and displayed within seconds, appropriate messages tell the operator what action to take on each form, data base update transactions are automatically generated for deferred updating, and custom programmed function keys allow operators to perform common operations with a single keystroke.

Integrity and security of the interactive system and the data base are maintained through a number of techniques: no dial-up access to the system is available, security passwords and eligibility criteria prevent unauthorized users from signing on the system and from using privileged routines, immediate validation and consistency checks are applied to the data entered at the terminal, and quality control operators systematically reprocess batches of completed work before overnight update of the data base.

2. Processing of administrative records--Administrative records from the Internal Revenue Service (IRS) and Social Security Administration (SSA) identify all firms that must be included in the economic censuses and provide basic data items such as the company name, mailing address, payroll, and employment. They also provide data on the industrial classification and legal form of organization. Proxy reports are derived from these records for most of the very small firms and other firms from which reports cannot be obtained. For those sent report forms, the data are used for comparison to reported information. These include such records as a master file of all legal business entities assigned an Employer Identification (EI) number for Federal tax reporting purposes; monthly changes, additions and deletions to the master file; quarterly payroll and employment records; and business income tax return information.

The dedicated computer has made it possible to process administrative records on a flow basis. Flow processing of forms, with an immediate match to administrative records, allows many of the processes that could not begin until the late fall when the entire file was processed sequentially to be completed as soon as the data from the questionnaire are keyed. Comparisons of historical and administrative records to reported data are made and anomalies resolved soon after receipt of the form.

3. Coverage and search processing -- Coverage and search processing routines allow the terminal operator to interactively identify owning or successor companies from a variety of information and update codes, addresses, and other records of companies and affiliated establishments. It also provides corrected information for labels for remailing report forms, and it generates codes during the processing that later generate labels for remail of report forms or letters requesting additional information.

4. Assignment of standard industrial classification (SIC) codes -- Interactive routines were developed to assist in the clerical assignment of SIC codes to individual report forms based on the respondent's self-description by capturing descriptive phrases describing types of primary

activity. Similar routines assign codes to unclassified companies based on the firm name.

The interactive SIC coding routines assume that the operator has a working knowledge of SIC coding and of the primary SIC divisions. The system requests the operator to enter a keyword that best describes the respondent's activity based on reported information or the company name. Using the keyword, the system retrieves and displays from an on-line dictionary all descriptions containing the keyword. The operator selects the description that appears to best describe the respondent's primary activity. If no selection is made, system algorithms decide whether to refer the report to an analyst or generate a letter to the respondent requesting additional classification information.

Compared to the manual coding methods used in the earlier censuses, the interactive system coders were roughly twice as productive, the learning time was substantially shortened, posting and separate data keying of transcription documents were eliminated, and summary and performance processing statistics were generated automatically. The error rates were roughly equal to those for manual coding. For the 1982 Economic and Agriculture Censuses, interactive computer assisted SIC coding was used to process almost a million reports and unclassified administrative records.

5. Microcomputers for the Census of Puerto Rico -- A microcomputer network was developed to process the census of Puerto Rico. These questionnaires are in Spanish and include a much wider range of industries than their counterpart forms for the mainland. Much of the delay in processing prior censuses of Puerto Rico was caused by the practice of modifying the processing procedures, the data entry programs, and the computer programs used in the main censuses to accommodate the listing book (used by the enumerators in the field for coverage of firms with no employees) and the unique forms used in that census. These modifications could not be made until each process for the main censuses was completed. Each counterpart phase for Puerto Rico was done last. By processing the Puerto Rico reports in an off-line operation, we were able to eliminate these built-in delays.

By moving the entire processing function to our temporary San Juan office, we were able to resolve problems on-site at an early stage. The microcomputer system performed the check-in function; schedules were edited on an interactive basis; administrative and historical records were checked by the computer; referrals were processed, and twice a week the edited records were sent on video cassette tape to Bureau headquarters. In prior censuses, the forms were sent to the Census Bureau's processing center in Jeffersonville, Indiana, and the processing was begun in mid-summer after the San Juan office was closed.

Planning

All phases of the censuses were reviewed for potential improvement during the planning phase, including the planning phase itself. Improvements in two planning operations stand out: (1) the development of a computerized time schedule and activity reporting system, and (2) simplification of the questionnaires and mailing packages.

1. Computerized time schedule and activity reporting system -- In previous censuses, certain problems were attributed to inadequate coordination of planning activities. It was recognized early that if our goals were to be met, no major operation could be delayed because of the failure to complete necessary procedures, specifications, computer programs, file preparation, or other operations in advance of need.

Various computerized reporting systems were reviewed. A system for monitoring projects through individual "nondependent" line activities was selected to assure that planning functions were completed in time. A system that shows the hierarchy of activities was established. Each activity record, representing a project, subproject, and line code contains information identifying the activity, type of operation, responsible division and individual, and planned and actual dates for both the starting and completion of the activity.

For the first eight months, the activity reporting system was maintained on a word processor, which limited the possibilities of special status reports. To rectify this shortcoming, the system was transferred to the Bureau's Univac 1100/83 computer. In addition to the information provided on the word processor, this permitted a variety of programs that provided additional information and reports.

2. Simplification of questionnaires and mailing package--Early in 1979 we began to analyze complaints received about the 1977 questionnaires. A major problem was one of perception. The questionnaire, four 10-1/2 x 17-inch pages, with an equal number of pages of file copies, gave the form a burdensome appearance. For the 1977 censuses, all forms, with the exception of the manufacturing short forms, were the same dimensions. Based on respondent comments, recipients, particularly small retail and service firms, often didn't bother to read the document to determine that relatively few data entries were required.

Another problem was that a single form, particularly in the service sector, often covered several diverse industries. Even though they were instructed to skip inquiries that were not applicable, respondents spent considerable time reading inquiries that did not apply to them.

A third problem was that all retail forms listed all major merchandise line categories. For example, shoe stores were asked about their sale of groceries, even though the results of previous censuses showed that this line almost never applied.

A fourth problem existed with the wholesale trade forms. The need to obtain data for both gross sales and commissions and the inabilty to obtain these two measures from administrative records, or even to determine which was available, has always precluded use of this source as a proxy for direct canvass. As a result, even the smallest firms in this census complete reports.

The 1982 censuses forms were tailored by industry more closely than ever before. As a result, only questions normally pertaining to the industry or industries receiving the form were included. This increased from 400 to 500 the number of different form types. For ease in photocopying by respondents, the forms were de-

signed letter size, legal size, or multiples of these sizes. Separate publicity brochures describing the purposes of the census were designed for each component census. A simplified transmittal letter also was developed.

A short form, which contained only limited inquiries, including gross sales and commissions, was developed for small wholesalers. For the first time, short forms also were developed for some industries in the construction census. The forms for the census of agriculture were redesigned and shortened to reflect regional crop patterns. And since its primary purpose was to obtain classification information from firms for which available information on economic activity was insufficient to determine the correct form type, the general purpose form was reduced to a single data item (in addition to several check boxes) from the multiple inquiries of prior censuses.

One measure of the effectiveness of these changes is the level of receipts by the due date. For the economic censuses, response at that time was 47 percent higher than in the 1977 censuses, which resulted in lower follow-up costs. The figures indicate that the increase was greatest in the trade and services sectors, where the forms were simplified most. A similar comparison was not possible for the census of agriculture, because reminder notices to stimulate response were mailed in advance of the due date in the previous census. Similar notices were not used for the 1982 censuses.

Another measure is the level of correspondence about census reports. Through mid-May, respondent-originated correspondence was less than one-half of the level in the last censuses. Congressional correspondence on behalf of constituents declined by more than two-thirds. The sharp reduction in the level of correspondence suggests that many respondents reacted by completing the report form rather than corresponding about it.

Processing of Report Forms

1. The check-in operation -- A potential bottleneck, particularly because of the uneven rate at which reports are returned, is the check-in operation. In previous economic censuses, and in agriculture censuses through 1974, reports were checked-in by data keying the identification number. Prompt recording of check-in actions is particularly important prior to mailing of follow-up letters to minimize complaints from those who have already submitted their reports. In 1978, the agriculture census successfully used supermarket-type bar coded labels to rapidly record receipt of reports and eliminate the keying of check-in actions. For the 1982 censuses, bar codes were used for the approximately seven million questionnaires in the combined program.

2. Sorting of report forms--Three high-speed six-pocket sorters with laser scanning capability were installed for initial sorting of the forms while still in the return window envelope as part of the check-in operation. In addition, the Bureau built two high-speed 24-pocket sorters, also with laser scanning capability. Schedules which required special attention because of their impact on the published data were identified for priority handling at time of check-in. This equipment was augmented with wand stations linked to microprocessors. In this operation, hand-held wands were passed over the bar code of reports that could not be read by the laser scanners. In addition to eliminating hand sorting, batching, and identifying of high priority reports, the system generated summary information, which eliminated manual counting for document control information. The system provided a valuable dividend in the mailing operation. The census law and the Bureau's uncompromising rules and procedures to ensure complete confidentiality of all information about individual firms require that all mailing packages for multiestablishment firms be verified to be certain that a form is not included in a package for the wrong company. This operation is particularly critical for firms with manufacturing plants, since their questionnaires may include data reported in the 1981 Annual Survey of Manufactures. By using the bar code and wand stations, this operation, which previously had been done clerically, was automated.

3. Data entry--One of the most costly and time consuming operations in census processing has always been transcribing information from report forms to machine language--the process of data entry or "keying." During peak processing of the previous censuses, more than 500 data keyers were used in a two-shift operation. Innovations for 1982 resulted in a dramatic increase in keyer productivity.

First, the keying methodology used in previous censuses consisted of using "fixed format" keying for items which appear on every form and "string" keying for items which varied from form to form. Numerous key strokes were eliminated in 1982 by using a system referred to as "packing the string" rather than the traditional method of associating a separate key code with each data item. Using the revised method, data items were inserted automatically in their proper positions at output, as part of the data entry machine function, rather than at input as part of the keying function. For many data items, this resulted in reductions of more than 50 percent in the number of key strokes required.

Other items on the form were examined to eliminate additional key strokes. By printing additional information in the label, we were able to avoid keying reported information, such as the county name, that was already in our records.

Second, training of keyers and monitoring of their progress was improved. Keyers were trained to specialize in only one type of form. If, because of lack of work, a keyer was moved, the move was to a form of equal or less difficulty, rather than to a more difficult form as was frequently the case in previous censuses.

The most important improvement, however, was managerial. In prior censuses, the Bureau operating units responsible for the component censuses (e.g., the retail census) prepared their own keying specifications. For the 1982 censuses, the subject divisions provided their requirements, while the detailed specifications for the keying programs were prepared centrally. A detailed time schedule was prepared, and all specifications were provided to the data entry programmers in the Bureau's Data Preparation Division (DPD) in Jeffersonville, Indiana in ad-

vance of need. This central planning of the data entry operation, which began with the forms design, made it possible for DPD management to complete testing and preparation of registers for hiring in advance of need and to complete training of the first group of keyers in advance of requirements.

As a result of these changes, 75 percent of the reports had been keyed and transmitted for computer edit by mid-May. The comparable percentage for the previous censuses was about 55 percent. This improvement was accompanied by lower error rates for almost all form types.

Preparation of Data for Publication

Improvements that reduced both costs and time were made in operations to prepare the tabulated totals for publication. Computer programs to prevent the disclosure of information about individual firms -- "disclosure analysis" -- were improved to eliminate almost all manual intervention. And major improvements were made to the computerized photocomposition system to produce publication "printer's copy" of the statistical tables.

1. Disclosure suppression--It is the policy of the Census Bureau to publish as much data as possible in each table without disclosing, directly or indirectly, data for individual firms. Publication cells for which data for individual firms would be revealed or could be closely approximated are always suppressed. To avoid disclosure of suppressed data indirectly through subtraction of published data from totals in the same or other tables, it is frequently necessary to suppress additional data cells -- an operation referred to as "complementary data suppression."

The 1977 Economic Censuses were the first for which a large-scale automated disclosure suppression system, utilizing a data base, was used. In prior censuses, only primary disclosures were identified and suppressed by computer; complementary suppressions were determined by analysts using hard-copy paper tabulation displays, a very time consuming operation. There were a number of shortcomings with the 1977 system. Substantial intervention by analysts was required. Because of the tremendous quantity of data that had to be processed, the program was very complex and costly to run.

To assure that confidential data are not inadvertently disclosed in other statistical tables, there are many considerations. The primary sensitive cells are identified by means of an n-k dominance rule. If n companies have values totaling more than k percent of the cell value, then that cell is considered sensitive and is suppressed. Additional cells must be suppressed to prevent the derivation of the primary suppressed data from published totals.

The major tabulation vectors for most publications are geography and SIC. They are usually hierarchical in structure. In a typical two-dimensional table, a comparison is made between SIC's and geography within the table. However, in analyzing the data and selecting complementary data suppressions there are many exceptions to be considered. Not all geographic levels and SIC's are considered equally important. For example, "catch all" or "other" categories usually range lowest on the priority scale. The most complex problems are geographic areas that are part of more than one state or county.

The large volume of data requires a partitioning structure to make the analysis manageable. These partitioned processing units, or LOGICAL-TABLES, consist of a tabulation breakdown along both vectors, geography and SIC. For example, any 2-digit SIC summed by its component 3-digits might be matched to a standard metropolitan statistical area summed by its component counties. The processing of these sets is done on a top-down basis to preserve the importance of data and to provide coverage for suppressions in total lines of the tabulations in its components.

A two-dimensional analyzer program is used to process the LOGICAL-TABLES. The analyzer takes advantage of knowledge of less important cells, such as in a catch-all category, and tries to funnel necessary suppression into these cells. It also attempts to preserve data in certain cells by avoiding their selection for complementary coverage. Within these constraints, the analyzer attempts to suppress as few cells as possible while suppressing the least total value of these cells as a secondary action. Detail data cells of a LOGICAL-TABLE are always preferred to a total cell.

The program size varies with the size of the LOGICAL-TABLE, e.g., a 3 x 5 matrix uses considerably less core space than a 20 x 120 matrix. Efficiencies in structuring these LOGICAL-TABLES and the use of specialized devices, rather than a generalized data base, to more effectively handle the input-output of the data have resulted in considerable reductions of real computer time for these programs. At the same time, the number of problem areas that cannot be resolved has been reduced; however, analysts still must resolve potential disclosure problems in these areas. Improvements in computer generated informational "flags" that explain the basis for suppressions greatly assist the analysts in completing the analysis of the disclosure suppression operation.

2. Computerized photocomposition system--For the 1977 Economic Censuses, a computerized photocomposition system, the Table Image Processing System (TIPS), was designed to produce publication quality statistical tables. Previously, standard format tables had been typed or generated by impact printers, with all titles, headers, vertical and horizontal rules, footnotes, symbols, and corrections added manually. Much of this was literally a "cut and paste" operation.

The 1977 system was a major innovation and greatly accelerated the production of publication copy. However, there were critical limitations. The system was limited to standard formats; and footnotes and corrections, including supplemental data suppressions by analysts, still had to be added manually. For 1982, major improvements were made to permit the photocomposition of any table format. The system also permits footnotes and corrections to be carried to the table by computer, virtually eliminating the need for manual intervention.

CONCLUSION AND PLANS FOR 1987

Concurrent processing of the economic and agriculture censuses was accomplished while re-

ducing respondent burden, and the results will be published earlier than ever before. Better tools, better methods, better design, and better controls were used in every stage of the operations, from initial planning through final publications. Objectives were met without compromising quality.

The 1982 censuses have provided a solid framework on which to build for the 1987 censuses. Several areas already have been identified where additional improvements can be made in the ways we collect, process, and review the data; disseminate the results; and manage these operations.

The content and design of the data collection forms will be further tailored to industry specialization; correspondence will be increasingly automated; and technological improvements will be made to the way we mail, open, sort, and check-in report forms. The Bureau is extending the use of computer assisted telephone interviewing (CATI), and we will be looking into this to improve data collection.

The capabilities of optical disk recording, image retrieval, and optical character recognition will be explored for potential improvements to our forms processing and data entry operations. We also will be looking at ways to integrate data entry processing more effectively with the mainframe computer operations to minimize forms handling and to streamline the complex edit and correction process.

The interactive processing capability developed for this census will be extended to additional areas, particularly data analysis. Computer terminals on each desk, on-line access to data, and additional software tools will allow analysts to analyze results better and faster.

The increased use of small computers is planned for decentralized processing and other special applications that need not or should not be on the mainframe. Management information, time and resource schedules, and additional outlying areas processing are initial candidates.

To avoid unnecessary duplication and to assure comprehensive solutions to common problems, research and planning for the censuses will be coordinated with that being done for other Bureau programs.

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to Mitchell Trager, Assistant Chief for Methodology and Program Development of the Census Bureau's Economic Surveys Division, who contributed greatly to this paper with his incisive review and comments.

FOOTNOTE

[1] Univac I was designed and built for the Census Bureau and was first used in 1951 for processing the 1950 Census of Population and Housing.