Survey Methodology, June 1989
Vol. 15, No. 1, pp. 119–127
Statistics Canada



# Automated Quality Assurance Procressing of Administrative Record Files

JAMES R. JONAS and PAUL S. HANCZARYK[1]

ABSTRACT

The Census Bureau makes extensive use of administrative records information in its various economic programs. Although the volume of records processed annually is vast, even larger numbers will be received during the census years. Census Bureau mainframe computers perform quality control (QC) tabulations on the data; however, since such a large number of QC tables are needed and resources for programming are limited and costly, a comprehensive mainframe QC system is difficult to attain. Add to this the sensitive nature of the data and the potentially very negative ramifications from erroneous data, and the need becomes quite apparent for a sophisticated quality assurance system on the microcomputer level. Such a system is being developed by the Economic Surveys Division and will be in place for the 1987 administrative records data files. The automated quality assurance system integrates micro and mainframe computer technology. Administrative records data are received weekly and processed initially through mainframe QC programs. The mainframe output is transferred to a microcomputer and formatted specifically for importation to a spreadsheet program. Systematic quality verification occurs within the spreadsheet structure, as data review, error detection, and report generation are accomplished automatically. As a result of shifting processes from mainframe to microcomputer environments, the system eases the burden on the programming staff, increases the flexibility of the analytical staff, and reduces processing costs on the mainframe and provides the comprehensive quality assurance component for administrative records.

KEY WORDS: Mainframe-microcomputer integration; Systematic data verification; Timeliness.

## 1. INTRODUCTION

The Bureau of the Census makes extensive use of administrative record information in our economic programs. The data originate from the business-related tax collection processes of the Internal Revenue Service (IRS) and, to a lesser extent, the Social Security Administration. During economic and agriculture censuses years, the volume of administrative record data received increases substantially. These data have enabled us to conduct economic and agriculture censuses on a timely and efficient basis and with a minimum of reporting burden on the business and farm communities. The success of our economic and agriculture programs depends to a great extent on the timeliness and quality of these administrative record files.

It is vital for Census Bureau operations to ensure the quality of all incoming data. As in past economic censuses, we have developed mainframe quality assurance programs for the administrative record data. However, since such a large number of these tables are needed and resources for programming are limited and costly, a comprehensive quality assurance system is difficult to attain entirely on the mainframe. Add to this the sensitive nature of these data and the potential ramifications of erroneous data, and the need for a more sophisticated quality assurance system becomes apparent. The Census Bureau has developed a comprehensive quality

[1] James R. Jonas and Paul S. Hanczaryk, Economic Surveys Division, U.S. Bureau of the Census, Washington, D.C., 20233, U.S.A.

assurance system that manages various phases of our administrative records review process. This automated system will allow us to perform more thorough quality assurance within the bounds of restrictive budgets and limited programming resources.

The automated quality assurance system integrates mainframe computer and microcomputer technology. The Census Bureau has established standards that delineate our fundamental requirements of the incoming administrative record data set. These standards are entered into a microcomputer system. After the mainframe quality assurance programs are run, the results are downloaded into the same microcomputer system. The reporting patterns of the actual administrative record data are then compared to the predetermined standards. Mechanical data verification occurs as data review, error detection, and report generation are accomplished automatically at the microcomputer level. As a result of shifting processes from mainframe to microcomputer environments, the system eases the burden on the programming staff, increases the flexibility of the analytical staff, and reduces the processing costs on the mainframe. Moreover, the system provides the quality assurance component needed for thorough and unerring review of administrative records. Although designed specifically for the IRS business income tax return files used in the censuses, it can and will be adapted to all incoming administrative record files after 1988.

## 2. OVERVIEW OF QUALITY ASSURANCE SYSTEM FROM A MANAGEMENT PERSPECTIVE

Administrative records play a major role at the Census Bureau, a role that has steadily grown in importance over time. The increasing need for more and better statistics, the need to compile those statistics with a minimum of burden on the private sector, and the need to use our available human and financial resources as efficiently as possible have all contributed to the importance of administrative records.

Over the past several years, the quality of the administrative records generally has been excellent. However, we did experience certain problems with the quality of the 1982 business income tax data from the IRS. The most detrimental problem was the inadequate quality of the principal industrial activity codes for sole proprietorships. As a result of this problem, the Census Bureau published only limited statistics for nonemployers in the 1982 Economic Censuses. If our quality assurance programs had been more sophisticated, the errors could have been identified earlier and the negative impact would have been minimized.

Heading into the 1987 Economic Censuses, it was determined that additional measures were needed to ensure the quality of administrative record data received from the IRS. An overall quality management system responsive to certain factors that have adversely affected past administrative data sets was necessary. The three major factors that have plagued us in the past are:

1. Vast amounts of administrative record data

   The IRS will provide us with selected business 1987 tax return data (received in 1988) for various legal forms of businesses, including corporations, S corporations, foreign corporations, partnerships, nonprofit organizations, and sole proprietorships. In total, the Census Bureau expects over 75 million tax return records in 1988. Table 1 details the approximate number of administrative records that will be used in the 1987 Economic and Agriculture Censuses for the various form types. Clearly, the number of data records received during census years is immense, but the complexity of the required quality assurance goes beyond sheer volume. A data record often contains several data items, each greatly increasing the detail of the individual records and the entire data files. Moreover, not all form types contain the same set of data items, nor do they have the same pattern of receipt. Consequently,

**Table 1**

The Approximate Number of Administrative Records Used in the 1987 Economic and Agriculture Censuses for the Various Form Types by Tax Year

| Type of Record | Number of Records | | |
|---|---|---|---|
| | 1985 | 1986 | 1987 |
| Business Income Tax Files | 2,617,000 | 20,051,000 | 30,881,000 |
| Form 1040, Schedule C | — | 11,750,000 | 12,500,000 |
| Form 1040, Schedule F | 2,450,000 | 2,450,000 | — |
| Form 1040, Schedule SE | — | — | 10,000,000 |
| Form 1120 | 42,000 | 2,550,000 | 2,650,000 |
| Form 1120-A | — | 200,000 | 210,000 |
| Form 1120F | — | 11,000 | 11,000 |
| Form 1120S | 17,000 | 900,000 | 950,000 |
| Form 1065 | 108,000 | 1,750,000 | 1,800,000 |
| Form 990 | — | 380,000 | 400,000 |
| Form 990-PF | — | 35,000 | 35,000 |
| Form 990-T | — | 25,000 | 25,000 |
| Form 1120S, Schedule K-1 | — | — | 700,000 |
| Form 1065, Schedule K-1 | — | — | 1,600,000 |
| Annual Tax Files | 41,950,000 | 43,500,000 | 45,050,000 |
| IRS Business Master File | 24,000,000 | 25,000,000 | 26,000,000 |
| IRS Payroll and Employment File | 17,000,000 | 17,500,000 | 18,000,000 |
| SSA Business Birth File | 950,000 | 1,000,000 | 1,050,000 |
| **Total** | **44,567,000** | **63,551,000** | **75,931,000** |

in addition to performing quality review for over 75 million individual records, the Census Bureau must also be concerned with assuring the quality of the various data items on those 75 million records.

Additionally, businesses file their tax returns with one of ten IRS centers. Each of the individual centers processes the returns, and the quality of data received from different service centers can vary. The Census Bureau reviews data at the service center level in response to such variation.

2. Restrictive budgets

Restrictive budgets are another major factor that contribute to the difficulty of assuring the quality of the administrative record data. In keeping with the overall governmental policy on spending, the Census Bureau is attempting to provide greater services at less cost. Workloads for programming staffs increase significantly during census years, yet the staffs do not expand proportionately. The quality assurance processing, which relies considerably on various computer resources, can be adversely affected. It is also important to note that most quality assurance processing is traditionally done at the mainframe computer levels. Use of the Census Bureau's mainframe computer is costly and becomes more so as increasingly larger data files are processed.

3. Lack of communication between agencies

Miscommunication or lack of communication between agencies has contributed to past administrative record problems. Clear lines of communication between the Census Bureau and the agency providing the data during all phases of the procurement process also are essential for assured data quality. The agencies first must agree upon the data files and the

specific data items that are needed and that can be provided. Certain data that the Census Bureau requests may not be available or in some cases affordable. Any discrepancies must be resolved in time to avoid delays, which could affect data utility. Moreover, the agencies must agree upon the expected quantity and quality of the administrative data. Requirements that quantify the Census Bureau's expectations of the incoming data should be established.

The development and implementation of the quality assurance system represent a comprehensive response to the administrative record data problems we encountered in the past. The system provides for the review of large and complex IRS data files, promotes frequent interagency communication, and identifies errors instantly. The major element of the quality assurance system is the mechanized data verification. Basically, the Census Bureau establishes standards that detail our fundamental requirements of the incoming IRS data. The reporting patterns of the actual data are compared to these standards, and systematic data verification occurs at the microcomputer level. The Census Bureau then prepares status reports indicating whether the data conforms to the standards.

Census Bureau staff members develop the standards far in advance of the actual receipt of the data. This gives the IRS ample opportunity to examine the requirements for reasonableness and request adjustments if necessary. The requirements are divided into timing standards and quality standards. The timing standards list the estimated total number of tax returns for the different types of businesses and the estimated number to be received by various dates. The quality standards detail the expected reporting patterns of specific data items.

The mechanized data verification technique simplifies our analytical review process. A series of results tables are created that compare the actual data to the expected standards. Discrepancy flags are set for those data components that do not meet the standards. This approach minimizes the risk of analytical omissions during the review process.

Status reports comparing the reporting patterns of actual data to the pre-determined standards are sent to the IRS monthly. These status reports are a subset of the comprehensive results tables, detailing only the basic requirements of the IRS data set. The status reports promote communication between the agencies. If data problems exist, they are illustrated in the report. Immediately, the Census Bureau and the IRS must decide upon any remedial action or recovery efforts necessary to prevent compromising the censuses. Timeliness is crucial because the IRS data tapes are not kept indefinitely. If errors are not identified early and remedial action is not implemented in time, recovery of the data may not be possible or may become extremely costly.

The quality assurance system is not designed to guarantee that administrative data problems will never occur. It does serve, however, to document our requirements formally so that the characteristics of the data set are not left to chance, and monitoring and early error identification are possible.

## 3. DETAILS OF AUTOMATION

Administrative record data files are received weekly and processed initially through mainframe quality assurance programs. The mainframe programs are prepared well before the administrative data files are received and generate the initial quality assurance tables that are fundamental to the entire review process. Traditionally, mainframe programmers were responsible for creating the entire data tables, which included data cells and the surrounding text (*i.e.*, headers and stubs). However, for the data table programs associated with the 1987 Economic Censuses, the two data table components are handled separately. Data tabulation is performed as usual at the mainframe level whereas table text is created at the microcomputer level by non programmers. A procedure has been developed that generalizes data tables for all administrative

**Table 2**

Weighted Distribution of Form 1040 Schedule C Records by Net Receipts Size Class by Service Center

| Service Center | Total | Net Receipts Size Class (000) | | | | |
|---|---|---|---|---|---|---|
| | | < 0 | Blank or 0 | 1— 2,499 | 2,500— 4,999 | 5,000— 9,999 |
| All Centers | 1,327,100 | 200 | 52,200 | 149,300 | 73,900 | 98,100 |
| Atlanta | 133,200 | 0 | 5,100 | 16,500 | 6,300 | 11,000 |
| Philadelphia | 132,100 | 100 | 4,200 | 11,300 | 5,300 | 9,600 |
| Austin | 147,600 | 0 | 6,300 | 20,900 | 9,900 | 12,900 |
| Cincinnati | 153,100 | 0 | 5,300 | 14,900 | 8,700 | 9,800 |
| Kansas City | 119,500 | 0 | 5,500 | 16,700 | 7,500 | 8,500 |
| Andover | 111,100 | 0 | 3,800 | 9,800 | 6,700 | 8,200 |
| Ogden | 162,300 | 0 | 7,500 | 20,200 | 7,900 | 11,600 |
| Brookhaven | 119,700 | 0 | 4,400 | 12,600 | 7,100 | 10,000 |
| Memphis | 111,900 | 100 | 4,700 | 14,700 | 6,700 | 8,600 |
| Fresno | 136,500 | 0 | 5,400 | 11,700 | 7,800 | 7,900 |
| Others | 100 | 0 | 0 | 0 | 0 | 0 |

| Service Center | Net Receipts Size Class (000) | | | | | |
|---|---|---|---|---|---|---|
| | 10,000— 24,999 | 25,000— 49,999 | 50,000— 99,999 | 100,000— 249,999 | 250,000— 499,999 | 500,000+ |
| All Centers | 168,600 | 185,500 | 225,100 | 243,400 | 87,400 | 43,400 |
| Atlanta | 17,000 | 19,800 | 22,200 | 22,200 | 8,400 | 4,700 |
| Philadelphia | 17,800 | 19,800 | 22,700 | 27,000 | 10,100 | 4,200 |
| Austin | 18,700 | 18,500 | 22,000 | 24,900 | 9,100 | 4,400 |
| Cincinnati | 20,500 | 20,700 | 27,300 | 30,500 | 9,600 | 5,800 |
| Kansas City | 16,200 | 15,900 | 20,700 | 18,300 | 6,400 | 3,800 |
| Andover | 13,600 | 16,700 | 19,500 | 20,000 | 8,800 | 4,000 |
| Ogden | 17,800 | 19,500 | 28,800 | 33,600 | 11,200 | 4,200 |
| Brookhaven | 16,400 | 19,700 | 20,400 | 19,400 | 6,400 | 3,300 |
| Memphis | 15,100 | 14,700 | 18,600 | 19,000 | 6,800 | 2,900 |
| Fresno | 15,500 | 20,200 | 22,900 | 28,400 | 10,600 | 6,100 |
| Others | 0 | 0 | 0 | 100 | 0 | 0 |

records files. This procedure has allowed the Census Bureau to design a microcomputer program that is capable of building table images for any administrative records file. Once built, the table images are uploaded to the mainframe and used by programmers to align data tabulation files. The job of programming the quality assurance tables is greatly simplified, as table image formation is handled by nonprogrammers, leaving mainframe programmers adequate time to concentrate their efforts solely on data tabulations. Table 2 illustrates one of the various mainframe tables that is produced for each of the different forms of organization. This table shows the weighted distribution of Form 1040, Schedule C records by service center by net receipts size class.

The mainframe computer performs only the basic data tabulations of the administrative records files (*i.e.*, generates current tables). The output from these mainframe quality assurance programs is downloaded to a microcomputer, and all remaining review operations are automated at the microcomputer level. The various operations performed on the microcomputer include calculating percentages used in the review of the current tables, producing

cumulative tables, performing key data item verification, and generating quality assurance status reports. Developing this systematic approach, using mostly micro-computer technology, has allowed greater flexibility of review as well as lessened the workload of mainframe programmers.

The mainframe quality assurance output is imported into a prestructured spreadsheet on the microcomputer. This spreadsheet also will contain the predetermined standards that outline the Census Bureau's expectations of the incoming data set. Automatically, a mechanical table review and data verification are performed; and inconsistencies between the actual data sets and the standards are identified within the results tables. The two major benefits of this data verification system are:

1. It enables us to easily spot problems in the data. Data components that do not meet the standards are flagged for analyst review. The possibility of overlooking errors in the administrative data is minimized.

2. It directs us to areas of the data that require further investigation. The results tables oftentimes lead us to problems even though the overall standards are met. For example, certain unexpected trends in the results report are reviewed in additional detail. In effect, the results tables enable us to concentrate on those areas that may contain problems. This may involve additional review at the service center level, or it may even require us to download records with these certain characteristics to the microcomputer. We then review these records on a manual basis in an effort to spot the problem.

As previously stated, the standards detail the basic data quality requirements that are essential to the 1987 Economic and Agriculture Censuses. This procedure of automatic quality verification (*i.e.,* comparing the incoming data to predetermined standards) allows us to determine immediately if the basic quality of the incoming data is acceptable.

After current cycle review and verification, cumulative tables are prepared on the microcomputer. This technique of producing cumulative tables on the microcomputer rather than the mainframe provides a more efficient use of our resources. First, it eliminates the need to retain cumulative files on the mainframe system, which reduces computer costs. In the past, these cumulative files were retained on the mainframe and added to each subsequent current cycle to form the next set of cumulative tables. Using microcomputers, simple formulas were established within the spreadsheet that created cumulative tables at virtually no cost. Secondly, the quality assurance tables for the cumulative portion do not require mainframe programming. A printout of the cumulative quality assurance tables are produced and retained for analysis and documentation purposes.

In addition to this comprehensive set of cumulative tables, we produce a set of results tables. As was the case with the current cycle, these results tables detail comparisons of certain key data items. Table 3 shows one of the many results tables that is produced for the cumulative quality assurance. This table details the actual number and percent of the weighted Form 1040, Schedule F records by service center, together with the expected percent. As can be seen, the cumulative data are reasonable and fall within the acceptable standards. If inconsistencies did exist, the applicable service center would have been flagged. The final component of the automated quality review process is the generation of a report detailing the status of the cumulative IRS data file. This report compares the overall quality of the data set to the expected quality indicated in the timing and quality standards. The reports are generated and provided to the IRS approximately monthly. As discussed earlier, the status reports capsulize the quality of the administrative data for representatives of both agencies, which promote frequent interagency communication.

## 4. RESULTS OF QUALITY ASSURANCE REVIEW

The timing and quality status reports can serve to alert both the Census Bureau and the IRS of data problems in their early stages and facilitate cooperative action by both agencies. In most of the cases, however, the timing and quality standards alert us of changes in respondent reporting patterns. These circumstances require no corrective action by the IRS, but they may have cost and processing implications for the Census Bureau in the 1987 Economic and Agriculture Censuses. Tables 4a and 4b illustrates this point well. Through late May 1987, the Census Bureau had received approximately 697,600 Form 1120 returns (*i.e.,* corporations) with a standard of 760,000 returns. The standard for the number of Form 1120 returns was not met. However, the shortfall in the number of Form 1120 returns was offset by an increase in the number of Form 1120S returns (*i.e.,* S corporations). The Census Bureau had received approximately 328,850 Form 1120S returns, far exceeding the standard of 225,000. The shift in the number of returns for these two types of corporations resulted from the perceived advantages in the new tax law associated with filing Form 1120S rather than Form 1120. Although this represented a legitimate shift in taxpayer reporting patterns that was not a data error, the information was pertinent to our processing. We are implementing a procedure for 1987 that will account for such a shift from corporations to S corporations. Table 5 illustrates one of the various tables from the quality portion of the report. As indicated, the quality of these data meets the standards for each of the basic data items. If an item had failed the standard, it would have been flagged for analyst research.

**Table 3**

Percent of Weighted 1986 Form 1040, Schedule F Records by Service Center

| Tax Year | Total Schedules | Service Centers | | | | |
|---|---|---|---|---|---|---|
| | | Atlanta | Philadelphia | Austin | Cincinnati | Kansas City |
| 1986 | | | | | | |
| Count | 2,087,200 | 176,700 | 71,600 | 374,900 | 262,100 | 358,600 |
| Percent | 100.0 | 8.5 | 3.4 | 18.0 | 12.6 | 17.2 |
| Expected | | | | | | |
| Percent | 100.0 | 8.5 | 3.0 | 18.5 | 11.5 | 17.5 |
| Expectation[1] Not Satisfied | | | | | | |

| Tax Year | Service Centers | | | | | |
|---|---|---|---|---|---|---|
| | Andover | Ogden | Brookhaven | Memphis | Fresno | Others |
| 1986 | | | | | | |
| Count | 118,800 | 343,200 | 40,300 | 288,100 | 52,500 | 400 |
| Percent | 5.7 | 16.4 | 1.9 | 13.8 | 2.5 | 0.0 |
| Expected | | | | | | |
| Percent | 5.5 | 16.5 | 2.0 | 14.0 | 2.5 | 0.0 |
| Expectation [1] Not Satisfied | | | | | | |

[1] Acceptance interval of + or − 2.0 percent.

**Table 4a**

The Weighted Number of 1986 Form 1120 Returns by Various Dates

| Date | Form 1120 Returns | | Requirement Not Satisfied |
|---|---|---|---|
| | Actual | Required | |
| Late March 1987 | 326,500 | 303,000 | |
| Late April 1987 | 697,600 | 760,000 | Not Satisfied |
| Late May 1987 | | 988,000 | |
| Late June 1987 | | 1,190,000 | |
| Late July 1987 | | 1,418,000 | |
| Late August 1987 | | 1,621,000 | |
| Late January 1988 | | 2,077,000 | |
| Late October 1988 | | 2,533,000 | |

**Table 4b**

The Weighted Number of 1986 Form 1120S Returns by Various Dates

| Date | Form 1120S Returns | | Requirement Not Satisfied |
|---|---|---|---|
| | Actual | Required | |
| Late March 1987 | 103,350 | 90,000 | |
| Late April 1987 | 328,850 | 225,000 | |
| Late May 1987 | | 292,000 | |
| Late June 1987 | | 352,000 | |
| Late July 1987 | | 420,000 | |
| Late August 1987 | | 480,000 | |
| Late January 1988 | | 615,000 | |
| Late October 1988 | | 750,000 | |

The automated quality assurance of administrative records files will be completely operational for the 1987 IRS data files. Prototypes of the system have been and are being used for the 1985 and 1986 IRS business income tax files. For both years the automated process and the entire quality assurance system have been instrumental in the successful procurement and review of the IRS data files received for the censuses.

The integration of both mainframe and microcomputer technology in the automated quality assurance system has allowed the Census Bureau to effectively and comprehensively assure the quality of the large data files provided by the IRS. In addition, mainframe computer programmer workloads have been and will continue to be lessened since much of the automation was designed and is controlled by nonprogramming staff and is implemented in a microcomputer environment. Mainframe computer resources are reduced and programming burden is lessened allowing programmers to concentrate their efforts on basic data tabulation. Also important, the automated system provides the flexibility of review for different levels of personnel. Managers can review the summarized timing and quality report and determine the status of the business income tax files quickly and efficiently. Subject-matter analysts will review the more comprehensive quality assurance reports that are produced weekly. As mentioned above, the quality assurance system will direct the analysts to the data elements that require further investigation.

**Table 5**
Data Element Reporting Patterns of Weighted 1986 Form 1120S Returns

| Data Elements | Percent of Form 1120S Returns | | Requirement Not Satisfied |
|---|---|---|---|
| | Actual | Required | |
| EIN | | | |
| Blanks, all zeros, or nonnumerics | 0.0 | Less than 1.0 | |
| Invalid IRD | 0.0 | Less than 1.0 | |
| PBA CODE | | | |
| Blanks or nonnumerics | 0.0 | Less than 6.0 | |
| Blanks, nonnumerics, unclassified, or invalid PBA codes | 11.5 | Less than 18.0 | |
| GROSS RECEIPTS OR SALES LESS RETURNS AND ALLOWANCES | | | |
| Blanks, all zeros, or nonnumerics | 20.9 | Less than 40.0 | |
| Of records with a positive numeric entry, the percent in various size ranges: | | | |
| - Less than $100,000 | 45.7 | 30.0 — 60.0 | |
| - Greater than or equal to $100,000 and less than $500,000 | 36.9 | 20.0 — 50.0 | |
| - Greater than or equal to $500,000 | 17.4 | 10.0 — 30.0 | |
| ACCOUNTING PERIOD | | | |
| Blanks, all zeros, or nonnumerics | 0.0 | Less than 1.0 | |

## 5. SUMMARY

The Census Bureau has designed an overall quality assurance system that is comprehensive and responsive to the potential problems and limiting factors of complete quality assurance. The system responds to the large volumes of IRS data by interacting with the IRS closely and promptly to ensure proper data procurement. The expected quality of these large data files is jointly determined and agreed upon with the IRS through the timing and quality standards and is verified by the automated QC process. Given this automated process, data verification can occur within the bounds of restrictive budgets and limited programming resources. Microcomputer technology has increased the role and flexibility of subject-matter analysts while lessening the burden of mainframe programmers. Communication with the IRS is frequent and productive, resulting in efficient procurement procedures and improved data quality awareness on the part of IRS and the Census Bureau as well. This collective response to past difficulties will ensure the Census Bureau of receiving the data necessary to conduct the 1987 Economic and Agriculture Censuses in the best manner possible.

## ACKNOWLEDGEMENTS

We would like to thank the referees for their comments and suggestions.