MAJOR ADMINISTRATIVE RECORD FILES: DOCUMENTATION AND POTENTIAL USES

Joseph J. Knott, Bureau of the Census

This paper describes the general properties of most of the major Federal administrative record files. An attempt is made to lay the groundwork and indeed begin the discussion, continued elsewhere in this session, of the current and potential statistical uses of these systems.

Organizationally, the paper is divided into four sections. The first section provides a list of the major administrative record files examined by the Subcommittee on the Statistical Uses of Administrative Records. Also described there is the survey instrument used to compile information for each of them. In the next section there is a brief summary of the survey results. The survey information forms a basis for initiating, in the last section, a discussion of some of the current and potential statistical uses of the major record systems studied.

SCOPE OF PAPER AND SURVEY CONDUCTED

Scope of Study.--In compiling a list of "administrative" record files that would be of greatest statistical interest, three criteria were employed:

1. Does the file have extensive coverage of a population (either individuals or businesses)?

2. Is the population covered by the administrative record set of statistical interest?

3. Is the file maintained by computer?

The systems chosen for examination under these criteria are shown in Figure 1. Information relating to individuals was sought from ten Federal agencies; some twenty-four administrative record files were involved in all. For businesses, the scope of the inquiry was restricted to nine major Federal systems in six agencies.

It should be noted that although the Subcommittee does not classify the decennial censuses of population as administrative data files, since their main purpose is statistical, they are nonetheless included to provide a basis for comparison with the other files on individuals. The Census Bureau's Standard Statistical Establishment List (SSEL) was also treated as "in scope" for comparison purposes, this time with business administrative record files.

Survey Conducted.--In late 1978, the Subcommittee conducted a survey of the administrative files listed in Figure 1. This survey was entitled "Statistical Use Survey of Records, Pertaining to Individuals, Individual Firms, and Employers Maintained and/or Mandated by the Federal Government."

For the survey, a questionnaire was designed and mailed to each agency maintaining one of the selected files. The principal purpose of the questionnaire was to document the data elements on each file that might be of statistical interest. It was not the intent of the survey to be comprehensive, but simply to provide a starting point for structuring further inquiries about the files. This survey collected data on both individual and business files by providing optional sections to be completed depending on the type of file being considered.

The survey consisted of only fifteen questions, but a number of the questions contained several parts. Respondents were asked to report the availability of documentation concerning the file, the information carried on the file, and the history of the file development and maintenance. For the most part, each agency made a serious effort to provide detailed responses to the questions.

SURVEY RESULTS

This section briefly summarizes the survey results. First, the files pertaining to individuals are considered, then those pertaining to businesses. Detailed tabulations from the survey are available from the author upon request. (A sample of those tabulations appears in the table at the end of this paper.)

Files Pertaining Mainly to Individuals.--Not unexpectedly, there are extensive differences among the administrative record files on individuals. Some of those which deserve special mention are the differences in coverage (or "universes") among the files, the degree of coded geographic information, the demographic items included and the reporting units used:

1. Universe.--In terms of coverage, the decennial census files are the most complete, followed by Social Security's Summary Earnings Files and the IRS Individual Master File. No other files have the same breadth of coverage as these. However, several other files do provide comprehensive coverage of important segments of the population. For example, the Health Insurance Master File--for the "65+" population; the Central Personnel Data File (OPM)--for Federal government workers; and the Military Personnel Data Files--for present and former Armed Forces members.

2. Geography.--Little coded geography exists on administrative files. Some contain a State code, but this was usually derived from the mailing address. The only exceptions appear to

be HCFA's Health Insurance Master File and the related SSA Master Beneficiary File, which contain a county code obtained by clerically coding the mailing address. By way of contrast, the Census geographic data are collected on a residence basis and are available to the block level.

This lack of detailed "residence geography" is a major problem in using administrative records to prepare small area statistics.

By using the mailing address, subcounty geography may be assigned with a Geographic Base File (GBF) developed for use in the 1970 or 1980 Census. However, this presents a number of problems. First, the mailing addresses are not always the usual place of residence. Second, GBF's do not exist for areas located outside the built up portion of SMSA's. Third, people living outside the city limits tend to report themselves as living in the city if they have a city post office address. Fourth, post office delivery or zip code areas do not conform with political boundaries. Also, the cost of assigning geography with a GBF system is high.

Another approach is to add a residence geographic code to the administrative file. This was done for the 1972 and 1975 Individual Master Files (IMF) so that IRS data could be used in preparing population and per capita total money income estimates for use in distributing General Revenue Sharing funds. The expense of this straightforward approach makes it unlikely that it will be widely implemented on other files.

3. Demographic Information.--By comparison with the Census data, all administrative files contain very limited demographic information. The Numerical Identification (SS-5) file does contain sex, date of birth, and race which have been transferred to the Summary Earnings Record and the Master Beneficiary Record. The personnel files also have some race information. However, other than this, there is very little demographic data present.

4. Reporting Unit.--The Census data are the only data organized into households and families. Tax returns and Social Security claims, however, can for some purposes be treated as approximations to family units. For the most part, however, the units are just individuals with no potential for structuring them into households.

One final point. The survey showed that all the administrative files for individuals are organized by social security number (SSN). This is distinct from the decennial census files which do not have the SSN recorded. By and large, the SSN is the major administrative identifier. Obviously, then, it is this variable which would have to be employed for linkages among the files--whether for statistical or operational purposes.

Files Pertaining Mainly to Businesses.--The employer identification number (EIN) is a major identifier on all the administrative record files, including even the Census' Standard Statistical Establishment List. Some other similarities and differences in the files are:

1. Universe.--The file with the largest coverage is SSA's Master Employer Name Directory with about 27 million records. However, this file is not current and contains inactive businesses. While not an administrative file, the Bureau of the Census' Standard Statistical Establishment List (SSEL) is the most comprehensive current list of businesses with the exception of the very small businesses. For these businesses, the IRS Business Master File is more complete. The Department of Agriculture's Producer Name and Address Master File, and their Economics, Statistics, and Cooperatives Service List Sampling Frame have extensive coverage of the farming sector.

2. Geography.--As with the individual record systems, there is not subcounty geographic data present on any of the business files with the exception of the Census Bureau's SSEL. For businesses, location may have different meanings. Most of the geography reported on these files is in terms of company headquarters and may not refer to the individual establishment. Consequently, a reporting of a major geographically dispersed company at its headquarters' location can introduce a significant error into the data.

3. Economic Data.--Number of employees, total payroll, and gross sales seem to be the most common economic items present on the files.

4. Reporting Unit.--The reporting unit of these files is mainly the Employer Identification Number (EIN) with the exception of the SSEL. This creates a problem in any statistical use of these files because some EIN's are only part of a company but an EIN may cover many establishments.

CURRENT AND POTENTIAL USES

The use of administrative records as a source of statistical information is not a new idea, but the last decade's extensive computerization of these files has fostered an increasing interest in the topic. In fact, there seems to have been a progression in the employment of administrative records for statistical purposes. Initially, with the establishment of an administrative records system, an agency prepared summaries of the data for guiding their operations and for policy decisions. This may be done with the full data set or a sample. Its purpose is primarily administrative, not statistical. Perhaps IRS is the best example. What started out as a mainly administrative effort has evolved into the current Statistics of Income program [1]. While administrative considerations are still important, the Statistics of Income sample is used extensively by researchers to study issues of general statistical and economic interest.

The administrative records systems were used very early in evaluation projects such as the evaluation of the 1950 Census income results using IRS and SSA data [2]. After each decennial population census since then, there have been attempts to understand and qualify any error or bias in the results by matching a small sample of census records to various administrative record sets such as IRS data [3], Medicare data [4], birth records [5], death records [6], and employment records [7].

These evaluation efforts may be characterized by the relatively small number of cases involved. This limit on size is the result of the objective of the project as well as cost considerations. Most evaluation projects involving these Federal files are aimed at National results only and do not attempt to measure differences at the State or even regional level. (This is changing, however; for the 1980 Census Evaluation, the matching will attempt to produce estimates at the State level [8].)

With the extensive computerization of these files in the 1960's, the possibilities for expanded statistical uses of administrative records became obvious. For example, IRS completed the computerization of the Individual Master File with the 1967 file. Also, over this same period, there was a great reduction in the cost of computer data processing and an increase in understanding how to process and control large data files, thus making the use of these administrative files feasible for statistical purposes.

These developments and potential uses of administrative records were understood and debated [9]. While that debate cannot be reviewed here, the outcome has been that no centralization of administrative records has taken place in the Federal government, but statistical uses of administrative records have continued. Some transfer of administrative records between agencies has been permitted, but each transfer has been justified and approved on a case-by-case basis [10]. Some people feel that this case-by-case approach has retarded the use of administrative records in developing useful statistical data, but this has never been fully documented.

In one sense, survey-and census-based data may be blamed for the slow development of administrative records-based data. Up until recently (and perhaps still), survey- and census-based data have had a real edge on administrative records in several areas. For example, if small area data are needed, the Census of Population and Housing provides small area data defined completely and in the "correct" geography (i.e., by residence). Administrative records-based data may be able to approximate the needed data, but not at the same level of accuracy. It is a question of trading-off precision and accuracy for currency. If the need is for National, regional, or even State data, surveys may be a more efficient way to obtain needed data than the development of an administrative records-based system.

However, with the need for small area data on a regular basis, the currency and small area advantages of administrative records may now outweigh the disadvantages of definitional problems and less accuracy. For example, with the passage of the State and Local Fiscal Assistance Act of 1972, the Bureau of the Census was asked to provide population and per capita total money income data for all 38,500 governmental units. The Bureau accomplished this by using an extract from the 1969 and 1972 entire IRS Individual Master File. This required IRS to collect and clerically code the residence address of all taxpayers on the 1972 IMF. The cost of the first set of estimates, including the IRS coding, was in excess of $5 million. This was the first administrative records-based project of this magnitude and demonstrated the expense and benefit of administrative records. It should also be noted that this successful application of administrative records used administrative records to measure change since the 1970 census [11]. In this way, the definitional problems were minimized.

With the expanded interest in administrative records, there is now taking place the needed experimentation and research to understand the particular idiosyncracies of these files. This will, hopefully, come to fruition in the 1980's with useful data in several areas. For example, migration rates by race can be computed by linking race from the SSA Summary Earnings File to the IRS data. This has been done on a sample basis and State estimates prepared [12]. It is expected that this work will continue.

By using tax returns (or W-2's) to establish a current residence, and the Form 941 to link an employer to an employee, and the Master Employer Name Directory (mainly SS-4) to define an employer's location, current journey-to-work estimates are possible. The Bureau of the Census and the Bureau of Economic Analysis (BEA) have done some work in this area, so far, however, without great success. The problems of multi-establishment employers, low quality geography coding of employers, etc., are major obstacles when trying to estimate the change in a particular journey-to-work flow.

Currently, the Census Bureau uses IRS adjusted gross income (AGI) and wages and salaries data to update the 1970 census per capita income estimates. By using the age, race, and sex data from the Social Security Administration, the IRS information could be adjusted for differential reporting by age, race, and sex. Updating the income size distribution with IRS data has long been considered desirable, but the inability to group IRS returns into families or households makes such updating difficult.

The need for targeted surveys and more sampling efficiency for small populations will continue to make administrative records important as a sampling frame. In the business files, the use of the business lists as sampling frames may be their single most important function, either to complete or to stratify a universe for sampling.

In summary, the statistical use of administrative records will continue to grow, but not easily. The use of administrative records data in preparing statistics must be preceded by a period of analysis and experimentation in order to understand the particular problems inherent in each administrative record system.

Figure 1.--Major Administrative Record Files Surveys by the Subcommittee on the Statistical Uses of Administrative Records

Agency	Administrative Record File
Part I.--Information on Individuals	
Bureau of the Census	1970 Census of Population 1980 Census of Population
Office of Personnel Management (OPM)	Central Personnel Data File Civil Service Annuity Roll
Department of Defense	Active Military Personnel Data File (Army, Navy, Air Force, and Marines) Military Retirement Compensation File (Army, Navy, Air Force, and Marines)
Department of Transportation	National Driver Register
Internal Revenue Service (IRS)	Individual Master File
Office of Education	Basic Education Opportunity Grant
Railroad Retirement Board	Research Master Beneficiary File Service and Compensation (SCORE) Railroad Retirement, Survivor, and Pensioner Benefit Payment File
Social Security Administration (SSA)	Summary Earnings Records Master Beneficiary Record Numerical Identification File (SS-5)
U.S. Coast Guard	Personnel Management-Information System Retired Officers Support system Retired Pay and Personnel System
Veterans Administration	Compensation and Pension Master Record Insurance (In-Force) Master Record File Education Master Record File Vocational Rehabilitation and Education Statistical File Insurance Awards Master Record File Education Master File
Part II.--Information on Businesses	
Bureau of the Census	Standard Statistical Establishment List (SSEL)
Bureau of Labor Statistics	Unemployment Insurance Address File
Department of Agriculture	Producer Name and Address Master File Economics, Statistics, and Cooperatives Service List Sampling Frame
National Center for Health Statistics	Master Facility Inventory
Internal Revenue Service	Business Master File (BMF) Exempt Organization Master File
Social Security Administration	Multi-Unit Code File (Establishment Reporting Plan File) Master Employer Name Directory

ACKNOWLEDGEMENTS

The author would like to thank several members of the Subcommittee on Statistical Uses of Administrative Data for their help in carrying out the survey described in this paper. Particular thanks are due to Dave Garnick and Maria Gonzalez for their leadership and coordination roles. Gratitude also needs to be expressed to Anne Richards and Joan Reynolds who typed the various drafts of this paper.

REFERENCES

[1] See, for example, Statistics of Income - 1977, Individual Income Tax Returns, 1979.

[2] Studies in Income and Wealth: An Appraisal of the 1950 Census Income Data, National Bureau of Economic Research, Studies in Income and Wealth, Vol. 23, 1958.

[3] Schneider, P. and Knott, J. Accuracy of Census Data as measured by the 1970 CPS-Census-IRS Matching Study, Proceedings American Statistical Association, Social Statistics Section, 1973, pp. 152-159.

U.S. Bureau of the Census, Evaluation and Research Program of the U.S. Population and Housing, 1960: Record Check of Accuracy of Income Reporting, Series ER60, No. 8, 1970.

[4] U.S. Bureau of the Census, 1970 Census of Population and Housing Evaluation and Research Program: the Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census, PHC(E)-7, 1973.

[5] U.S. Bureau of the Census, Infant enumeration study: 1950 completeness of enumeration of infants related to: residence, race, birth month, age and education of mother, occupa- of father, Procedural studies of the 1950 Census, No. 1, 1963.

U.S. Bureau of the Census, 1970 Census of Population and Housing Evaluation and Research Program: Test of Birth Registration Completeness 1964 to 1968, PHC(E)-2, 1973.

[6] Kitagawa, E.M., and Hauser, P.M., Differential Mortality in the United States: A Study in Socioeconomic Epidemiology, Harvard University, Cambridge, 1973.

[7] U.S. Bureau of the Census, Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Employer Record Check, Series ER60, No. 6, 1965.

[8] Bateman, D.V. and Cowan, C.D., Plans for 1980 Census Coverage Evaluation, 1979 Proceedings American Statistical Association, Section on Survey Research Methods.

[9] See, for example, the Report of the President's Commission on Federal Statistics, 1971; especially Hansen, M.H., The Role and Feasibility of a National Data Bank, based on Matched Records and Interviews, Vol. 2, pp. 1-63.

[10] Kilss, B. and Scheuren, F., (with F. Aziz and L. DelBene), The 1973 CPS-IRS-SSA Exact Match Study: Past, Present and Future, Policy Analysis with Social Security Research Files, U.S. Social Security Administration, 1979, pp. 163-194.

[11] Fay, R. and Herriot, R., Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, J. Amer. Stat. Assn., 1979, pp. 269-277.

[12] Word, D.L., Population Estimates by Race, for States: July 1, 1973 and 1975, Current Population Reports, Special Studies, Series P-23, No. 67, 1978.

SAMPLE

Table 1. -- MAJOR ADMINISTRATIVE RECORD SYSTEMS PERTAINING TO INDIVIDUALS

Type of Information	Bureau of the Census			
	1970 Census		1980 Census	
	100%	Sample	100%	Sample
File Organization	Residence Code	Residence Code	Residence Code	Residence Code
Number of Records (Approximate)	200,000,000	70,000,000	220,000,000	75,000,000
Data:				
Primary Type of Unit	Individual by Household	Individual by Household	Individual by Household	Individual by Household
Name	No	No	No	No
Address	No	No	No	No
Coded Geography	Residence	Residence	Residence	Residence
Race	Yes	Yes	Yes	Yes
Spanish	No	Yes	Yes	Yes
Date of Birth or Age	Quarter and Year	Quarter and Year	Quarter and Year	Quarter and Year
Sex	Yes	Yes	Yes	Yes
Marital Status	Yes	Yes	Yes	Yes
Income	No	Yes	No	Yes
Employer	No	No	No	No
Occupation	No	Yes	No	Yes
Education	No	Yes	No	Yes
Year Computer File First Established	1970	1971	1980	1981

SAMPLE

Table 2. -- MAJOR ADMINISTRATIVE RECORD SYSTEMS PERTAINING TO BUSINESSES

Type of Information	Bureau of the Census	Bureau of Labor Statistics	Department of Agriculture	
	Standard Statistical Establishment List	Unemployment Insurance Address File	Producer Name and Address Master File	ESCS List Sampling Frame
File Organization	Company/Establishment/EIN	UI Number	Not Available	Various (File is maintained in each State)
No. of Businesses on File	150,000/5,500,000/4,000,000	4,000,000	5,000,000 individuals and businesses engaged in farming	2,700,000 for operators
Availability of File Documentation	On request	On request	On request	Not available outside agency
Type of Documentation	General Description	Layout and Tech. Description	Not Available	None
Data:				
Name	Yes	Yes	Yes	Name of operator (may be person or business)
Address	Yes	Yes	Yes	Yes
Location Code	Yes	Yes	State and County	Yes
Date of Determination of Number of Employees	March 12 Pay Period	First Quarter	No	No
Total Payroll	Yes, annually and quarterly	Yes, quarterly	No	No
Primary Industry & Coding System	Yes, 4-digit SIC Minimum	Yes, 4-digit SIC	No	Farming
Gross Sales and Receipts	Yes	No	No	Yes, Usually
Product Description	No	No	No	Type of Farm
Form Used	Various sources including SSA's SS-4, IRS forms, and the Bureau of the Census' Company Organization Survey	State Unemployment Insurance	CCC-181 Master Name and Address List	Various
Computerized "Paper System"	Yes	Not Available	Yes	No
Year Created	1972	Not Available	1973	1978
Date Expanded or Changed	Not Available	Not Available	Not Available	Still being constructed
Purpose of SSN or EIN on File	EIN for identification	EIN for identification	EIN/SSN for identification	EIN/SSN for identification