Statistical Policy Working Paper 6

Report on Statistical Uses Of Administrative Records

1980
U.S. DEPARTMENT OF COMMERCE
Office of Federal Statistical Policy and Standards

Statistical Policy

Working Paper

Report on

Statistical Uses Of

Administrative Records

Prepared by

Subcommittee on Statistical Uses of Administrative Records

Federal Committee on Statistical Methodology

U.S. DEPARTMENT OF COMMERCE
Philip M. Klutznick, Secretary
Luther H. Hodges, Jr., Deputy Secretary
Courtenay M. Slater, Chief Economist

Office of Federal Statistical Policy and Standards
Joseph W. Duncan, Director

Issued: December 1980

Statistical Policy Working Papers are a series of technical documents prepared under the auspices of the Office of Federal Statistical Policy and Standards. These documents are the product of working groups or task forces, as noted in the Preface to each report.

These Statistical Policy Working Papers are published for the purpose of encouraging further discussion of the technical issues and to stimulate policy actions which flow from the technical findings and recommendations. Readers of Statistical Policy Working Papers are encouraged to communicate directly with the Office of Federal Statistical Policy and Standards With additional views, suggestions, or technical concerns.

Office of
Federal Statistical
Policy and Standards

W. Duncan
Director

For sale by the Superintendent of Documents,

U.S. Government Printing Office
Washington, D.C. 20402

Office of Federal Statistical

Policy and Standards

Joseph W. Duncan, Director

Katherine K. Wallman, Deputy Director, Social Statistics
Gaylord E. Worden, Deputy Director, Economic Statistics
Maria E. Gonzalez, Chairperson,
Committee on Statistical Methodology

Preface

This working paper was by the members of the Subcommittee on Statistical Uses of Administrative Records, Committee on Statistical Methodology. The Subcommittee was chaired by Daniel H. Garnick, Bureau of Economic Analysis, Department of Commerce. The members of the subcommittee are the authors of this report, their names are listed below.

The first portion of this report provides a review of major administrative report files pertaining to individuals and to businesses. Major statistical uses of administrative records are outlined, including: (1) direct use of the records to obtain statistics and to supplement existing data via expanding coverage or content; and (2) technical uses of the data for constructing sampling frames, quality control, improving on procedures, and data evaluation. New developments in data from business establishment reporting and a number of potential uses of administrative records

for data linkage are described. Technical problems in the statistical use of administrative records, including coverage, comparability, error and timing of data are discussed. the final section of the report covers various in accessing administrative records for statistical purposes.

While much statistical use of administrative records is currently made in Federal agencies, this report is intended to inform managerial and technical staffs of the vast potential as well as difficulties entailed in augmenting current uses of administrative records for statistical purposes. The Office of Statistical Policy and Standards hopes to organize, with the help of Subcommittee members, seminars with Federal employers to disseminate the findings of this report. The implementation of the recommendations in report will be explored by the Office of Statistical Policy and Standards.

Members of the Subcommittee on Statistical
Uses Of Administrative Records.
(June 1980)

Daniel H. Garnick* (Chair)

Bureau of Economic Analysis (Commerce)

Lois Alexander
Social Security Administration (HHS)

Paul A. Armknecht
Bureau of Labor Statistics (Labor)

David V. Bateman
Bureau of the Census (Commerce)

Lawrence A. Blum
Bureau of the Census (Commerce)

Warren L. Buckler
Social Security Administration (HHS)

David W. Cartwright
Bureau of Economic Analysis (Commerce)

John DiPaolo
Internal Revenue Service (Treasury)

Maria E. Gonzalez* (ex officio)
Office of Federal Statistical Policy & Standards (Commerce)

John A. Gorman
Bureau of Economic Analysis (Commerce)

David A. Hirshberg
Small Business Administration

Beth A. Kilss
Social Security Administration (HHS)

J. Knott
Bureau of the Census (Commerce)

Bruce Levine
Bureau of Economic Analysis (Commerce)

Nash J. Monsour
Bureau of the Census (Commerce)

Allan Olson
Economic Development Administration (Commerce)

Elizabeth H. Queen
Bureau of Economic Analysis (Commerce)

Vernon Renshaw

Bureau of Economic Analysis (Commerce)

Fritz J. Scheuren*

Social Security Administration (HHS)

Daniel F. Skelly

Internal Revenue Service (Treasury)

Hyman Steinberg

U.S. Postal Service

Additional Contributors to the Report on Statistical Uses of Administrative Records

Jeanne E. Griffith

Office of Statistical Policy and Standards (Commerce)

Daniel Kasprzyk

Assistant Secretary for Planning and Evaluation (HHS)

Susan Miskura

Bureau of the Census (Commerce)

* Member, Committee on Statistical Methodology

Acknowledgments

The body of this report is the collective effort of the Subcommittee on Statistical Uses of Administrative Records. Although the subcommittee members reviewed and commented on all parts of this report, specific individuals were responsible for preparing the various sections. In the case of Chapter VI, the subcommittee benefitted from the expertise and contribution of

several additional persons in preparing the case studies. The authors of the chapters appear below:

Chapter	Authors
I	Daniel Garnick, Maria Gonzalez, Vernon Renshaw, Lois Alexander, David Hirschberg, Fritz Schuren
II	Vernon Renshaw, David Hirschberg, Daniel Garnick
III	Joseph Knott, Lawrence Blum, Waken Buckler, Vernon Renshaw, Fritz Scheuren
IV	Vernon Renshaw, David Cartwright, Nash Monsour, Lawrence Blum, John Gorman, Daniel Skelly, John DiPaolo, Warren Buckler, Elizabeth Queen
V	Lawrence Blum, Paul Armknecht, Warren Buckler, David Cartwright, Vernon Renshaw
VI	Fritz Scheuren, Beth Kilss, Jeanne Griffith, Daniel Kasprzyk, David Bateman, Sue Miskura, Maria Gonzalez
VII	David Cartwright, Vernon Renshaw, Bruce Levine, Warren Buckler, Fritz Scheuren
VIII	Lois Alexander

Maria Gonzalez worked with the subcommittee throughout its two-year study. Members of the Federal Committee on Statistical Methodology and the Office of Statistical Policy and Standards provided additional assistance and encouragement. Critical reviews of earlier draft versions by Thomas Jabine, Barbara Bailar, and Tore Dalenius were particularly helpful in the development of this report.

Discussion by Richard Ruggles on papers by Daniel Garnick and Joseph Knott, David Cartwright and Paul Armknecht, David Hirschberg and Vernon Renshaw, and Lois Alexander at the Statistical Uses of Administrative Records Session of the 1979 American Statistical meetings aided in sharpening the focus of this report.

Others who contributed to the work of the Subcommittee include: Yoshio Akiyama, Leroy Bailey, Robert Berney, J. Robert Brown, Morris M. Kleiner, Lillian Madow, Harriet Orcutt, and Max Shor.

Members of the Federal Committee on

Statistical Methodology

(June 1980)

Maria Elena Gonzalez (Chair)

Office of Federal Statistical Policy and Standards (Commerce)

Barbara A. Bailar

Bureau of the Census

Norman D. Beller

Economics, Statistics, and Cooperatives Service (Agriculture)

Barbara A. Boyes

Bureau of Statistics

Edwin J. Coleman

Bureau of Economic Analysis (Commerce)

John E. Cremeans

Bureau of Economic Analysis (Commerce)

Marie D. Eldridge

National Center for Education Statistics (Education)

Daniel H. Garnick

Bureau of Economic Analysis (Commerce)

Thomas B. Jabine

Energy Information Administration (Energy)

Charles D. Jones

Bureau of the Census (Commerce)

William E. Kibler

Economics, Statistics, and Cooperatives Service (Agriculture)

Alfred D. McKeon

Bureau of Labor Statistics (Labor)

Raymond C. Sansing

Internal Revenue Service (Treasury)

Fritz J. Scheuren

Social Security Administration (HHS)

Lincoln E. Moses

Energy Information Administration (Energy)

Monroe G. Sirken

National Center for Health Statistics (HHS)

Wray Smith

Office of the Assistant Secretary for Planning and Evaluation (HHS)

Thomas G. Staples

Social Security Administration (HHS)

Table of Contents

page

Preface i

Acknowledgments iii

List of Figures ix

List of Tables x

Abbreviations xi

Chapter I. Findings and Recommendations 1

A. Statistical Standards 1

B. Access 2

C. Other Government-Wide Program Coordination and Support 3

Chapter II. Introduction and Summary 5

A. Introduction 5

B. Summary 6

1. Chapter III 6

2. Chapter IV 7

3. Chapter V 8

4. Chapter VI 8

5. Chapter VII 8

6. VIII 9

Chapter III. Major Administrative Record Files 11

A. Scope of Study and Survey Conducted 11

1. Scope of Study 11

2. Survey Conducted 12

B. Survey Results 12

1. Files Pertaining Mainly to Individuals 12

a. Universe 12

b. Geographic Information 17

c. Demographic Information 17

d. Reporting Unit 17

2. Files Pertaining Mainly to Businesses 18

a. Universe 18

b. Geographic Information 18

c. Economic Data 18

d. Reporting Unit 18

C. Continuous Work History Files 18

D. The Evolution of Statistical Uses of Administrative Records 19

E. Appendix III.1 The Survey Questionnaire 20

F. Appendix III.2 The CWHS Data System 23

1. Data Sources 23

2. Processing Procedures - Administrative Records 23

3. Processing Procedures - Statistical Records 24

4. Sample Design 24

5. Data Files 25

a. One percent Sample Annual Employee-Employer (Ee-Er) File 25

b. One percent Sample Annual Self-Employed (SE) File 25

page

c. One percent Sample Longitudinal Employee-Employer Data (LEED) File 25

d. One percent 1937 to Date CWHS File 26

c. One-Tenth of One percent 1937 to Date CWHS File 26

Chapter IV. Major Statistical Uses of Administrative 27

A. Defining Administrative and Using Them Statistically 27

B. Internal Revenue Service 28

C. Social Security Administration 29

D. Bureau of Economic Analysis 29

E. Census Bureau 32

1. Economic Censuses 32

2. Census Of Agriculture 33

3. Survey of Minority-Owned Businesses (SMOBE) 33

4. Current Economic Indicators 33

5. The Standard Statistical Establishment List 34

F. The Small Business Administration 34

G. Appendix IV.1 Data from IRS and SSA 35

1. Data from IRS 35

2. Data from SSA 39

Chapter V. Developments in Data from Business Establishment Reporting 43

A. Standard Statistical Establishment list 43

1. File Construction 44

2. Multiestablishment Firms 44

3. Single Establishment Firms 44

4. File Maintenance 45

5. Confidentiality 45

B. W-2 and W-3 Records 45

C. Unemployment Insurance System 47

1. Master List of Employers 47

2. Employers' Quarterly Tax Report 47

3. Individual Wage Records 49

4. Improving Data Quality 49

Chapter VI. Potential Uses of Administrative Records for Data

linkages: Selected Case Studies 51

A. Introduction 51

B. Case Study 1: Linked Administrative Statistical Sample (LASS) Project 51

1. Background and Initial Project Goals 52

a. LASS Data Elements 52

b. LASS Goals 53

2. Pilot Activities and Feasibility Issues 53

a. Resolving Privacy Concerns 53

b. Examining SSA-NCHS Death Reporting Differences 54

c. Adding Data From Death Certificates to the CWHS 54

d. Usability of IRS Occupation Information 54

c. Upgrading CWHS Industry and Place of Work Data 55

f. Evaluating W-2 Residence Data 55

3. Operational Implementation Issues 56

4. References 56

C. Case Study 2: The Use of Administrative Records in the Survey of Income and Program Participation 57

1. Objectives and Description 58

page

a. Site Research 58

b. 1978 Panel 59

c. 1979 Panel 61

2. Major Difficulties 61

3. Uses of Administrative Files 62

4. Quality of Results 63

5. Bibliography 63

D. Case Study 3: Use of IRS/SSA/HCFA Administrative Files for 1980 Census Coverage Evaluation 64

1. Introduction 64

2. Objectives of the Program to Estimate the Census

Undercount 64

3. Matching Techniques 65

a. Matching of Survey Housing Unit and Person Records to Census Records 66

b. Matching of CPS and Census Enumerated Housing Unit and Person Records to Administrative File Records 66

4. Administrative matching 66

5. Research Conducted for Match Study 67

a. 1978 CPS/IRS Match Study 67

b. IRS Census Match Study (Involving Richmond Virginia and Southwest Colorado Dress Rehearsal Censuses) 68

6. Estimation 68

7. Anticipated Cost and Timing of Administrative Match Study 70

8. References 70

E. Case Study 4: Record Linkage in the Nonhousehold 70

1. Introduction 70

2. Results from the Travis County, Texas and Camden New

Jersey Pretest 71

3. Plans for the 1980 Census Nonhousehold Sources 75

4. Summary and Future Considerations 76

5. Sources of Further Information 77

6. References 77

7. Appendix-Matching Instructions 78

F. Concluding Comments 78

Chapter VII. Technical Problems in the Statistical Use of Administrative Records 81

A. Coverage 81

B. Comparability 83

C. Reporting and Processing Errors 85

1. Reporting Problems 86

2. Processing Problems 86

3. Extent of Errors 97

4. Related Problems with Other Data 88

5. Errors in Other Information 98

D. Problems with Timing of the Data 89

E. Conclusion 89

Chapter VIII: Legal Issues in the Statistical Use of Administrative Records 91

A. Legal and Administrative System 91

1. Factors Precipitating the Shift Toward Greater Statistical Use of Administrative Records 91

page

2. Concept of Functional Separation 92

3. A Language Framework for Legal Issues 93

4. Options: Legislative Approaches to Functional Separation 95

B. Dynamics of Functional Separation 96

1. Dimensions and Characteristics of the Legal Framework 96

a. Disclosure Within the Agency, a Broader View 96

b. Disclosure to Agency Contractors 97

c. Disclosure Among Federal Agencies 98

d. Use By Non-Statisticians of Statistical Files Compiled From Administrative Source Records 99

2. A Closer Look At Some Federal Statutes Affecting Statistical Use of Administrative Records and Protection of Statistical Records from Nonstatistical Use 100

C. Summary and Directions for the Future 102

D. Notes and References 102

References 104

List of Figures

Figure		page
III.1	Major Administrative Files Surveyed by the Subcommittee on the Statistical Uses of Administrative	11
V.1	Forms W-2 and W-3	46
V.2	Statistical Uses of Unemployment Insurance Administrative Records from Establishments	48
VI.4.1	Nonhousehold Sources Worksheet to Search Census Records for Selected Person: 1976 Census of Travis County,Texas	72
VI.4.2	Nonhousehold Sources Census Record Search and Telephone Follow-up Verification Record: 1976 Census of Camden, New Jersey	75
VI.4.3	Nonhousehold Sources Record: 20th Decennial Census-1980	76

List of Tables

Figure		Page
III.1	Major administrative Record Systems Pertaining to Individuals	12
III.2	Major administrative Record Systems Pertaining to Businesses	15
IV.1	National Income and Product Account Components Based on Administrative Records	30
IV.2	Input-Output Account Industry Estimates Based on Administrative Records	30
IV.3	Balance of Payment Account Components Based on Administrative Records	31
IV.4	National Income and Product Account Components Based on Current Surveys Using Administrative Record Based Sampling Frames	31
VI.2.1	Distribution of Site Research Sample Households by Sample Frame and Questionnaire Type	58
VI.2.2	Distribution of Site Research Adult Respondents by Sample Frame and Questionnaire Type	59

VI.2.3 A Sampling of AFDC Matching Results in the Site Research Survey 60

VI.2.4 SSI Match Results for the 1978 Panel 60

VI.3.1 Forming a Dual-System Estimate for One of the 61 Divisions 69

VI.4.1 Camden Match Results 73

VI.4.2 Cross Tabulation of Age Reported on Drivers Licenses and Census Questionnaire (Camden, New Jersey) 74

VII.1 Comparison of Employment Estimates: CWHS, Census, UI, and CBP 84

Abbreviations

AFDC Aid to Families with Dependent Children

BEA	Bureau of Economic Analysis
BEOG	Basic Education Opportunity Grant
BLS	Bureau of Labor Statistics
BMF	Business Master File (of IRS)
CAB	Civil Aeronautics Board
CBP	County Business Patterns
CofC	Comptroller of the Currency
CES	Current Employment Statistics
CETA	Comprehensive Employment and Training Act
CPS	Current Population Survey
CWBH	Current Wage and Benefit History
CWHS	Continuous Work History Sample
ED	Enumeration District
EEOC	Equal Employment Opportunity Commission
EI(N)	Employer Identification (Number)
ERP	Establishment Reporting Plan
FAA	Federal Aviation Administration
FCC	Federal Communications Commission
FDIC	Federal Deposit Insurance Corporation
FICA	Federal Insurance Contributions Act

FOIA	Freedom of Information Act
FPC	Federal Power Commission
FRB	Federal Reserve Board
FTC	Federal Trade Commission
GAO	General Accounting Office
GBF	Geographic Base File
HCFA	Health Care Financing Administration
HHS	Department of Health and Human Services
HEW	(Department of) Health, Education, and Welfare
ICC	Interstate Commerce Commission
IMF	Individual Master File (of IRS)
I-O	Input-Output
IRS	Internal Revenue Service
ISDP	Income Survey Development Program
LASS	Linked Administrative Statistical Sample
LTS	Labor Turnover Statistics
NCEUS	National Commission on Employment and Unemployment Statistics
NCHS	National Center for Health Statistics
NCI	National Cancer Institute

NIPA	National Income and Product Accounts
OASDI	Old Age, Survivors, and Disability Insurance
OES	Occupation Employment Statistics
OFSPS	Office of Federal Statistical Policy and Standards
OMB	Office of Management and Budget

OPM	Office Of Personnel Management
ORS	Office of Research and Statistics (of SSA)
OSHS	Occupation Safety and Health Statistics
PES	Post Census Enumeration Survey
PPSC	Privacy Protection Study Commission
REA	Rural Electrification Administration
RFP	Request for Proposal
SBA	Small Business Administration
SER	Summary Earnings Record

SESA	State Employment Security Agency
SIC	Standard Industrial Classification
SIPP	Survey of Income and Program Participation
SMD	Statistical Methods Division (of Census)
SMOBE	Survey of Minority Business Enterprises
SMSA	Standard Metropolitan Statistical Area
SOI	Statistics of Income
SSA	Social Security Administration
SSEL	Standard Statistical Establishment List
SSI	Supplemental Security Income
SSN	Social Security Number
SSR	Supplemental Security Record
SUAR	Statistical Uses of Administrative Records
TCMP	Taxpayer Compliance Measurement Program
UI	Unemployment Insurance
USDA	United States Department of Agriculture

CHAPTER I

Findings and Recommendations

Statistical use of administrative records grew rapidly during the 1970's, in large part as a response to legislative requirements for timely data to use in the distribution of Federal funds to State and local governments. The principal reason for increasing reliance on administrative records for statistical data is the availability of administrative records which can be used to obtain small area data at minimal cost and without increasing respondent burden. And cost is likely to be an increasingly important factor in the statistical use of administrative records in the 1980's. Although statistical use of administrative records is growing, many unanswered questions remain concerning the quality of statistics derived from administrative records. From a statistical point of view, the standards of quality and consistency in administrative data collection and processing programs are frequently inadequate.

Difficulties in accessing administrative records, moreover, often inhibit the efficient joint use of particular administrative record sets with other administrative and statistical records in meeting statistical needs. Improved statistics from administrative records will require modification in data collection and processing procedures, modification of laws and administrative procedures relating to access to records, and increased resources for evaluating and upgrading the quality of administrative records for statistical use. While the costs of improving administrative records for statistical applications can be significant, they will often be substantially less than alternatives requiring expanded censuses and surveys. And in many instances both administrative and statistical programs could benefit from reduced respondent burdens and data processing costs obtainable by applying more efficient statistical tools in the collection and use of administrative records.

To solve problem impeding efficient statistical use of administrative records, coordinated treatment of a variety of interagency issues is needed to serve as a counterweight to the decentralized operations of Federal information collection

programs. In addressing these issues, the Subcommittee on Statistical Uses of Administrative Records has divided its recommendations into dim sections concerned with:

A. Identifying and formulating solutions for common problems related to statistical standards for administrative information programs.

B. Identifying and meeting various problems related to access to administrative record systems.

C. Identifying collection programs and research activities requiring government-wide coordination and support.

Individual recommendations are in some cases accompanied by examples of subcommittee findings which illustrate the need for the recommendation.

A. Statistical Standards

There is a need for greater standardization in the procedures for collecting and presenting data based on administrative records in order to provide a basis for reducing duplicate collection efforts and improving the quality and consistency of the

information that is collected.

Recommendation 1. Common identifiers should be used whenever possible in collecting information Pertaining to the sow individuals or organizations.

The capability for linking information from a variety of sources is central in making efficient statistical use of administrative records. This capability depends on both appropriate access to administrative records (see Section B) and consistency among administrative and statistical agencies in procedures for identifying respondents or reporting units. The subcommittee noted, for example, that household surveys could be used more effectively in conjunction with administrative records if social security numbers and related identifying information were collected in selected surveys. This would permit linking detailed socioeconomic information from surveys with longitudinal records from administrative sources concerned, for example, with employment or medical histories. Such linkages are performed in various areas of social research including specialized fields such as epidemiology. In business data collection programs, employer

identification numbers should be supplemented with a common set of identifiers for the individual establishments of large businesses. Selected administrative record data for multi-establishment businesses could then be linked more readily to economic census and survey data for purposes of improving geographical and industrial analysis of economic activity

Recommendation 2. The quality of administrative records to be used for statistical purposes should be evaluated systematically to determine the appropriateness of the records for the proposed use.

The quality of administrative record files, including such factors as the type and quality of identification on the file and the completeness, definitional suitability, and quality of individual or organizational characteristics on the file. will

determine the appropriateness of the use of the files for particular statistical applications. For example, in matching applications the completeness of the coverage of the administrative record files and the accuracy of identifiers will determine whether a high match rate will be achieved. Similarly, in such applications as the distribution of Federal funds to State and local governments. completeness and accuracy of administrative records, will determine the extent to which estimates derived from these records may serve as complements as well as substitutes for census and survey data.

Recommendation 3. Consistent procedures should be used in administrative and statistical data collection efforts for defining reporting units, identifying and coding reporting unit characteristics, and developing standards for data tabulation.

When common reporting units are not appropriate there should still be efforts to ensure that the more detailed reporting unit breakdowns of one program can be readily combined into more aggregative units used in other programs. The subcommittee noted, for example, a lack of congruity in the definition of companies

filing corporate income tax returns and companies reporting for statistical Purposes to the Census Bureau. The subcommittee also found a particularly serious problem of inconsistency between "establishment" reporting plans associated with administrative programs and the definitions of establishments of multiunit companies used in the Census Bureau's Standard Statistical Establishment List. The Social Security payroll tax program, for example, involves a voluntary establishment reporting plan with company self-identification of reporting units on a basis differing from SSEL definitions. The need for consistent reporting requirements that eliminate duplicate and other unnecessary reporting is highlighted by the fact that the compliance of large companies with the SSA establishment reporting plan and other voluntary statistical programs has been deteriorating in recent years.

Problems of inadequate procedures for coding reporting unit characteristics have been emphasized by the subcommittee in such areas as geographic coding and the industrial coding of business establishments. Reliable and detailed geographic coding in administrative record systems, in particular. has become increasingly important as administrative records have received

wider application in preparing statistics for use in distributing Federal funds to State and local governments. For many purposes geographic coding is required at the municipal level, but substate coding in administrative record systems tends to be restricted to county identifiers. The lack of current economic information by municipality has hindered effective planning and economic policy making at the Federal as well as State and local level. For business reporting systems, the SSEL coding system can provide a basis for obtaining consistency in both geographic and industrial coding.

The need for consistent standards for data tabulation have recently been highlighted by efforts to assemble a data base for analyzing small business policy issues. These efforts have been hampered by inconsistencies among various administrative and statistical programs in the ways in which data are identified and tabulated by size of business.

B. Access

A central issue related to meeting the differing requirements

of data for administrative vs. statistical applications efficiently involves the problem of obtaining an appropriate balance between the need to access individual records and the right to privacy as well as consideration of confidentiality of responding persons and businesses. Resolution of this issue requires that distinctions be made both in terms of the uses to be made of records and the types of reporting units and information involved.

Recommendation 4. Natural persons should be distinguished from organizations and other entities when developing standards and practices of record confidentiality.

The need for confidentiality is not the same for businesses and other organizations as for natural persons. Often,, the need for access to selected information pertaining to businesses requires interagency transfer of information about organizations. The subcommittee has found, for example, instances in which Federal a#coca purchase privately produced lists of businesses containing generally available information, such as name and address of the businesses, because access to more complete and reliable lists such as the Census Bureau's SSEL has been excessively restricted. The subcommittee is not persuaded that these restrictions are

reasonable or necessary.

Recommendation 5. Legislation and administrative procedures should be modified to make comprehensive Federal lists of businesses and organizations, such as the

Census Bureau's Standard Statistical Establishment List and SSA's employer listing, more readily available for statistical uses.

Legislation has been drafted to make the SSEL available to Federal agencies for statistical purposes. Passage of the proposed legislation could aid in reducing the duplication and costs, and the attendant differences in definition and coverage resulting when independently developed lists are maintained. SSA's listing of employers is compiled from the applications for employer numbers required of employers of workers covered by Social Security, now

virtually the entire workforce. Availability of this list as a statistical sampling frame has been closed by application of the Tax Reform Act of 1976.

Recommendation 6. For natural persons. the principles of "functional separation" developed by the Privacy Protection Study Commission, the White House Privacy Initiative, and the President's Statistical Reorganization Project should be applied in distinguishing records to be used for administrative (and enforcement) purposes from records to be used for statistical purposes.

Functional separation will establish two discrete categories of information according to the statistical or administrative and enforcement functions to which the information is assigned. The separate category of statistical information- can be freely used and transferred with individual identifiers intact for statistical purposes. Between the two categories, information that can be uniquely associated with subject individuals flows only one way, into the statistical category. The flow from the statistical category into other uses must be in a form or under conditions that prevent unique association. When administrative records are the initial information source, the resultant copies or extracts which

have been incorporated into statistical files may not be subsequently used in individually identifiable form for administrative or enforcement purposes.'

Recommendation 7. Particular legal and administrative barriers to access to administrative records for statistical use should be identified and eliminated for records pertaining to both natural persons and organizations.

The subcommittee, for example. has found limitations on access to IRS data imposed under Section 6103 of the Tax Reform Act of 1976 to be excessively restrictive to statistical uses of the data. In this connection it can be noted that the Internal Revenue Service has denied other Federal agencies access to Taxpayer Compliance Measurement Program data files for 1976 and subsequent years. In addition, the Tax Reform Act has prevented the Social Security Administration from supplying the Bureau of Economic Analysis with post- 1975 Continuous Work History Sample Files needed to continue a long-standing cooperative program to use and improve this important statistical data base.

C. Other Government-Wide Program Coordination and Support

In order to maximize the usefulness of administrative record systems, it will be necessary to identify on a government-wide basis those data collection programs, as well as research initiatives, which need interagency support. Further the needs of data users should be considered in designing statistical series based on administrative records.

Recommendation 8. Procedures for planning and setting budget priorities should be developed to ensure that agency and program-specific budget allocations are responsive to those interagency data needs that are met most effectively through the specific programs under review.

Many administrative programs are not explicitly budgeted for supplying those general-purpose statistical needs which could be met efficiently through statistical use of administrative records. The subcommittee has found, for example, that geographic and industrial data quality in the Social Security Administration's Continuous Work History Sample has been declining because the data

have few applications for internal SSA programs and therefore receive low priority in the agency budgeting process. Geographic and industrial data from the CWHS, however, are very important for outside data users. And they will become even more important if administrative records are called on to play a central role in providing intercensal estimates. In planning alternatives to a mid-decade census there should be careful cost-benefit analysis of different approaches involving various combinations of survey and administrative record data sources.

Recommendation 9. As recommended by the President's Statistical Reorganization Project, efficient statistical tools should be applied in information collection programs extending well beyond the confines of the principal statistical agencies.

Statistics can contribute techniques for improving design of forms. both to improve quality of response on administrative forms, and to improve the multi-purpose utility of the information provided. Development and extension of such statistical techniques as scientific sampling. record matching, and synthetic estimation can be used effectively to economize on the amount of information that needs to be collected, thereby reducing paperwork burdens and

budgetary costs associated with administrative as well as statistical data collection programs.

Many administrative record data collection programs have lagged well behind the "state of the art" in the application of statistical tools, and modernization of programs is badly needed.

Recommendation 10. To obtain statistical data. increased use should be made of matches between sample surveys and administrative files. Samples based on linkageS among administrative record systems also should be encouraged for statistical purposes.

The subcommittee has investigated the statistical uses of linking of administrative record files with sample survey data. as well as with samples from other administrative records. The subcommittee endorses the use of matching to obtain statistical data based on the combination of administrative records and sample

surveys. The analytic potential of obtaining expanded. more detailed data bases through successful matching is sufficiently great that complicated procedures are often worth the effort. However, for each specific program proposing to use linkage s to obtain statistical data. it is necessary to examine the costs and benefits to the program to determine whether the match should be performed.

The case studies in Chapter VI illustrate potential uses of administrative records for important statistical programs'. each case study has specific goals, applications, and advantages. Mc combined use of administrative record files and sample survey data for linkage programs may be effective for a variety of masons. including that: (1) respondent burden may be reduced while estimates of subpopulation characteristics are improved and data accuracy is assessed (see SIPP case study), (2) data which are difficult for a survey respondent to provide may be obtained from administrative record files (see LASS case study). (3) improved counts of population from the 1980 Census may be obtained in a cost-effective manner (see Nonhousehold Sources Program case study), and (4) estimates of coverage of population for States and

selected subgroups of the population based on the 1980 Census my be obtained (see case study on IRS/SSA/HCFA matched with CPS and Census).

Recommendation 11. The provision o f services to users should be recognized as a statistical program function to optimize the availability of statistical information in Federal. State and local government and in the private sector, and to give the Federal system the benefit of feedback from users in planning statistical programs based on administrative records.

A major obstacle to encouraging statistical use of administrative records is the lack of knowledge (both inside and outside the Federal Government) about the information in these records and their coverage and quality. The American Statistics Index provides a comprehensive list of published statistics from administrative and survey sources, but information on the quality and availability of unpublished data, particularly from administrative records, is seriously deficient. Centralized information is needed to make existing data more readily accessible to potential users and to help in identifying unnecessary duplication in data collection programs. Promising recent

initiatives in this area include a Small Business Administration program to document all Federal reporting requirements placed on businesses and a National Center for Health Statistics program to establish a clearinghouse for data relating to environmental health hazards. In addition, the proposed Paperwork Reduction Act of 1980 (H.R. 6410) provides for establishing a Federal Information Locator System, as recommended by the Commission of Federal Paperwork.

CHAPTER II

Introduction and Summary

A. Introduction

The Federal Statistical System is under pressure to respond simultaneously to a growing demand for statistical data and a growing demand for reductions in the "paper blizzard" generated by Government requests for information from individuals and businesses. These demands will necessarily conflict unless the efficiency of current programs can be improved. Responsiveness to both demands will require reduced duplication among Government information collection programs combined with more intensive utilization of existing administrative information sources in meeting statistical data needs. The latter requirement will involve bringing together information collected in numerous different Government administrative programs in ways that make possible their combined use for statistical analysis. As stated by Edgar Dunn (1965, P. 5) in a review of the Ruggles' Committee proposal for a national data center.

The central problem of data use is one of associating numerical records. No number conveys any information by itself. It acquires meaning and significance only when compared with other numbers. The greatest deficiency of the existing Federal Statistical System is its failure to provide access to data in a way that permits the association of the elements of data sets in

order to identify and measure the interrelationship among interdependent activities.

As Dunn further notes (1965, Summary, p. 2) problems of access and record association are particularly serious in the case of statistical use of administrative records because: "Many of the most useful records are produced as a by-product of administrative or regulatory procedures by agencies that do not recognize a general-purpose statistical service function as an important part of their mission."

The association or merger of administrative records from a variety of sources is important for statistical applications because: (1) populations of statistical interest do not always correspond closely to populations covered in individual administrative record systems; and (2) individual administrative record files often identify relatively few of those characteristics and attributes of the members of a population that social scientists and policy analysts consider to be important in meeting their statistical needs. Merging individual administrative record sets with other administrative and statistical data sources can help to alleviate the deficiencies of many individual administra-

tive sources; but record merging is often difficult--particularly when the records are collected and maintained by separate agencies. Provisions for protecting the confidentiality of records pertaining to identifiable individuals or businesses often preclude interagency transfer of such records for statistical applications. And even when access to the records needed for merging can be arranged, differences in the ways different agencies identify individual reporting units, and/or inconsistencies in the ways agencies collect, process, and maintain information about reporting units, can preclude successful data matching and merging operations (see Chapter VI).

Although difficult problems remain to be solved, statistical uses of administrative records have been increasing and will continue to increase because of high data collection costs and heavy respondent burdens associated with censuses and surveys. Many important statistical needs cannot be adequately met by a system involving censuses, carried out every 5 or 10 years, combined with intercensal surveys which provide national data. And the extra costs of moving to more frequent censuses and/or larger sample surveys which might provide small area data are high both in

terms of direct government expenditure and response burden. The projected high cost to the government was an important factor in the recent decision to disallow further planning funds for the 1985 mid-decade census.

The most striking illustrations of the need to make improved statistical use of administrative records arise in cases involving the use of socioeconomic data to distribute Federal funds to State and local areas. For example, in reviewing alternatives for meeting the legislative mandate to produce current local-area unemployment estimates for use in allocating funds under the Comprehensive Employment and Training Act, the National Commission on Employment and Unemployment Statistics (1 979, p. 253) has estimated that it would cost about $2.3 billion annually to expand the Current Population Survey to provide monthly unemployment estimates for the over 4,000 geographic areas potentially eligible for CETA funding. As important as the high money costs involved in obtain-

ing frequent small-area data by survey techniques is the substantial increase in response burdens associated with greatly expanded data collection efforts.

For example, another alternative considered by the NCEUS was improving the handbook method (called 70-step method) based on unemployment insurance records.

Not only is them pressure for statisticians to increase their use of administrative records in developing general-purpose statistics, but statisticians also have a strong interest in supporting efforts to reduce the duplication and improve the efficiency of administrative as well as statistical information collection efforts. Direct reporting for statistical purposes accounts for a very small proportion of the overall Federal reporting burden; major reductions in overall paperwork burdens must be achieved through improvements in nonstatistical arm. At the same time; however, statistical programs could be more adversely affected than other programs because statistical programs

tend to be more often viewed as optional than administrative record systems and, therefore. more dependent on the voluntary cooperation of the public in obtaining responses to information requests.

As the following statement from the President's Statistical Reorganization Project's "Issues and Options" paper (1978, p. 7-1) indicates, there is a growing recognition of the importance of applying statistical tools to more general problems of information collection in order to reduce reporting burdens:

> The tools used by statistical agencies (sampling, quality control, intensive analysis of existing data, etc.) are near the roots of reporting requirements, and the use of appropriate tools reduces reporting burden. It is in this sense that. from the point of view of response burden, the use of appropriate statistical techniques is of major importance and should extend well beyond any formal definition of the Federal Statistical System.

The statistical system, however, cannot hope to dominate Government information collection activities; There must be a genuine effort to cooperate with administrators in nonstatistical programs in order to achieve mutual goals of efficient information collection.

Statisticians must attempt to understand the needs and constraints facing program administrator and statistical budgets should bear a fair share of the costs of collecting and processing administrative records in ways that permit efficient use for statistical purposes. Much must be learned and many difficult problems confronted if progress is to be made in the statistical use of administrative records and in improving the overall efficiency of Government information collection and use, With the hope of contributing to progress in this area, this report attempts to: (1) identify major administrative data files with significant potential for general-purpose statistical applications; (2) indicate various kinds of statistical uses of administrative records which are being made or considered; (3) identify major technical and institutional or legal problems which are impeding effective statistical use of administrative records; and (4) suggest possible approaches to improving information collection and statistical use of administrative records.

The Subcommittee on Statistical Uses of Administrative Records has not attempted to provide comprehensive documentation of administrative record systems and their uses. The report instead

reflects largely the areas of interest and expertise of Subcommittee members. Important areas such as energy and environmental statistics are not covered at all, and very little attention is given to records generated by the complex array of Government regulatory agencies. There is, however, relatively intensive coverage of administrative data from programs of the internal Revenue Service and Social Security Administration, and from related administrative programs that collect important social and economic information from individuals and businesses.

B. Summary

Chapter III of the report presents the results of a survey conducted by the Subcommittee to obtain documentation of major administrative record data files maintained by selected Federal agencies. Chapter IV presents a description of statistical applications of administrative records in selected agencies. The following three chapters (V-VII) illustrate, largely by means of case studies, specific approaches to statistical use of

administrative records and problems encountered in such approaches. Chapter VIII reviews legal considerations, particularly those related to restricted access to records, that influence the statistical use of administrative records.

1. Chapter III-Major Administrative Files

This chapter summarizes the characteristics of major computerized administrative record files that are maintained or mandated by the Federal Government and contain statistically useful information pertaining to (I) individuals or (2) businesses. The information contained in the administrative files for individuals is compared to the information on individuals collected in decennial censuses; and the information contained in the administrative files for businesses is compared to the information contained on the Census Bureau's Standard Statistical Establishment List (which is itself assembled from a combination of administrative and survey data sources). The chap-

ter also contains a description of the Social Security Administration's Continuous Work History Sample which is a set of statistical files of individual worker records assembled using several SSA business and individual administrative record files.

Compared with the decennial census, most administrative record files for individuals contain relatively little information on population characteristics and/or cover only a limited segment of the population. In addition, the, census usually provides more reliable and detailed geographic information than administrative files; and at best, administrative records can provide only tough approximations to such census reporting units as the family and household. On the other hand, many administrative files provide data at much more frequent intervals than the decennial census, and the presence of social security numbers on most administrative files opens the possibility of linking files over time (longitudinally) or merging information from more than one

administrative file in order to increase the cove rage of individuals and/or the number of characteristics identified for particular individuals. The absence of SSN's in census records generally makes it difficult to integrate information from censuses with information from administrative records.

Administrative record coverage of businesses is complete than is true for individuals. In fact, administrative lists of businesses provide the basis for conducting statistical censuses and surveys. For the most part, however, administrative records do not maintain separate information for the different establishments of a single legal business entity, even though the business may operate in several different geographic areas and/or industrial categories. The Census Bureau does collect information for individual establishments; and the SSEL, therefore, contains a larger list of reporting units than most administrative files. While most administrative business files do not contain the establishment detail necessary for developing reliable geographic and industrial data, the SSA and Unemployment Insurance payroll tax programs do involve reports breaking out county level "establish-ment" detail. Unfortunately, however, the reporting units in these

programs are not consistent with the establishment concept used in the SSEL, and there is currently no satisfactory basis for coordinating the reporting of similar information (or resolving data discrepancies) among the three systems.

CWHS data files provide information on the demographic characteristics (sex, age, and race) of. workers along with longitudinal information on their employment and earnings patterns. The CWHS program illustrates the potential statistical advantages of administrative records for longitudinal analysis and for linking together information about individuals and businesses.

2. Chapter IV-Major Statistical Uses of Administrative Records

This chapter illustrates statistical uses of administrative records with reference to the programs of selected Federal agencies, particularly programs of the Social Security Administration, the Internal Revenue Service, the Bureau of Economic Analysis, the Census Bureau, and the Small Business Administration. The SSA and IRS programs involve the development of general-purpose statistics by statistical divisions of agencies

that collect large amounts of information from individuals and businesses in the course of their administrative responsibilities. The programs illustrate the large quantity and variety of administrative data collected as well as the limitations of incomplete population coverage and lack of information on important population characteristics that plague statistical use of administrative records.

The BEA programs illustrate the use of a wide variety of administrative data (obtained from many agencies) for estimating data series within the context of a systematic economic accounting framework. Administrative data are used in conjunction with census and survey data (also generally obtained from other agencies); and there are substantial variations among the administrative data series in the extent to which they involve concepts and measurement procedures that "fit" well with the concepts involved in the design of the accounting framework and with concepts underlying the census and survey data used.

Census Bureau programs illustrate a wide variety of applications of administrative records for both individuals and businesses. For example, records obtained from administrative agencies are used in developing intercensal population and related

estimates, as a substitute for censuses in the collection of economic data from many small businesses, in the development and maintenance of sampling frames for surveys, and in the evaluation of the completeness and, reliability of information collected in censuses and surveys. Again there are substantial variations in the extent to which administrative record concepts match desired statistical concepts. A few census programs. primarily in the area of economic statistics. art discussed in more detail than other programs covered in Chapter IV. These more detailed examples illustrate the substantial cost savings as well as limitations associated with the statistical use of administrative records.

The SBA involvement in the statistical use of administrative records stems largely from a recently initiated project to develop a small business data base in conjunction with the 1980 White House Conference on Small Business. In part because of concerns over reporting burdens, small businesses have been exempted from or

covered on a very small sample basis, in most economic censuses and surveys. Therefore. a small business data base must rely heavily on administrative records. SBA efforts to develop such a data base illustrate many of the problems that are often encountered in gaining access to administrative records and adapting them for statistical analysis.

3. Chapter V-Developments in Data from Business Establishment Reporting

This chapter contains case studies of three important and related statistical programs that are currently evolving based in large part on developments in administrative record systems-(1) the Census Bureau's SSEL program; (2) SSA's program for adapting its CWHS data program to a new system of annual employer reports of worker wages on forms W-2 and W-3; and (3) the Bureau of Labor Statistics' program for developing work force statistics in

connection with the UI payroll tax program. These programs produce both complementary and overlapping statistical products in the area of work force statistics; and they illustrate not only the importance and potential of administrative records for developing work force data, but they also illustrate some important problems in the area of establishment reporting by multiestablishment businesses and in the area of coordinating similar data collection efforts in different agencies. The Census Bureau program employs the most satisfactory concept of establishment from a statistical point of view, but the Census work force data assembled in connection with the SSEL cannot match the frequency and timeliness of BLS data based on the UI system, nor can the SSEL-based data provide the information on demographic characteristics of workers available from the SSA system. And the different establishment reporting plans of the three data systems combined with difficulties of interagency transfers of records (for example, the current restrictions on access to the SSEL) have severely limited the scope for coordinating data collection and development efforts in the three programs.

4. Chapter VI--Potential Uses of Administrative Records for Data Linkages: Selected Case Studies

This chapter involves four case studies that illustrate the potential and the problems associated with record linkages as a means of improving and extending the use of. administrative records in developing primary data and in evaluating census and survey data--(1) the "Linked Administrative Statistical Sample Project" (2) the "Use of Administrative Records in the Survey of Income and Program Participation," (3) the "Use of IRS/SSA/HCFA Administrative Files for 1980 Census Coverage Evaluation," and (4) "Record Linkage in the Nonhousehold Sources Program." In contrast to Chapter V, where the difficulties of coordinating and linking business establishment records among programs was highlighted, Chapter VI is concerned with linkages involving records for individuals.

The LASS project involves efforts to link records from a variety of administrative record sources in order to develop a general-purpose statistical sample file that will be suited for mortality research. The sampling procedures will conform closely to those involved in the CWHS in order to facilitate longitudinal data analysis, but CWHS records will be supplemented with records from IRS and the

National Center for Health Statistics. The project illustrates the substantial potential for combining complementary data through interagency linkage of administrative record files. But the project also illustrates significant technical problems and problems of access restriction that need to be resolved in linking data files prepared in different agencies.

The SIPP case study illustrates the importance of administrative records in efforts to alleviate substantial survey biases in coverage and income reporting for low-income groups (participating in various income maintenance programs) and administrative record importance as a source of income data to evaluate the reliability with which selected types of income are reported in surveys.

The third and fourth case studies are both associated with efforts to evaluate and improve the 1980 Census of Population and Housing. The IRS/SSA/HCFA files will be used primarily in efforts to evaluate the extent of Census undercoverage, while the Nonhousehold Sources Program will be concerned with improving population coverage in selected areas of anticipated high undercount. The latter program involves, in addition lo the use of

Federal agency records, the use of such State and local administrative records as drivers' license records. Both projects demonstrate the potential of administrative records to identify individuals who are missed in censuses and surveys. The projects also illustrate; however, the difficulties and high costs of linking administrative records to census records (which contain no social security number) and the difficulty of determining the extent to which particular groups are not covered in either census or administrative record sources.

5. Chapter VII-Technical Problems in the Statistical Use of Administrative Records

This chapter illustrates technical problems encountered in making statistical use of administrative records that arise or are exacerbated because of limited statistical control in administrative record systems over such factors as population coverage,, definitions and comparability of information concepts among programs, and reporting and

processing procedures. The CWHS data program is used as the principal source of illustrations, in part because the CWHS program involves the use of files containing information about businesses as well as individuals, and perhaps more importantly because it illustrates well the problems that can arise when important statistical aspects of the reporting and processing of records we largely outside the control of statisticians responsible for making statistical use of the records. In particular them is evidence of significant and increasing numbers of geograPhic coding errors in the CWHS that have resulted from low priority attached by SSA administrators to the statistical problem of obtaining reliable geographic reports and ensuring accurate coding and processing of geographic information in employer payroll reports to SSA.

6. Chapter VIII: Legal Issues in the Statistical Use of Administrative Records.

This chapter illustrates legal and related institutional barriers which inhibit the interagency access to records that is needed for improving the efficiency and effectiveness of statistical use of administrative records. Emphasis is placed on problems which arise because of a failure of existing confidentiality laws to make an adequate functional distinction between statistical and administrative processes which use records about individuals.

The basis for interagency transfer of administrative records is often found in a logic that imposes regular Procedures or conditions for expanding the scope 'of administrative actions or decisions which can be based on the. particular content of records about an individual. Such a logic is generally irrelevant with respect to legitimate statistical processes which, in contrast to administrative uses, merely produce relationships and summaries of data, and do not involve any direct Government action against (or in favor of) the individual as a consequence of information in records pertaining to that individual.

Clearly not all statistical performance is functionally divorced from administrative processes: program integrity and quality assurance are functions which may explicitly---and quite properly-rely on applied statistical techniques to identify individual cases for administrative action. Such functions are within the reasonable expectations of program participants, and do not rely, moreover, on collection of information from volunteers, with assurances of confidential treatment. In contrast, there are particular statistical activities or collections of data whose existence and rationale for compiling and making interagency transfer Of data is limited by the degree to which statisticians can fulfill a legal or ethical duty to protect the confidentiality of individual information.

Statistical uses in this latter category need to be separated out as discrete functional uses, and be governed by different rules and standards from those which govern administrative and compliance uses. Proposals for functional separation" of statistical from administrative uses argue for separating these statistical records about identifiable individuals from the decision/action stream, and permitting the statistical results to be available to adminis-

trators only in summary or other unidentifiable form. Functional separation would allow summaries, of course, to be used administratively in ways which my result indirectly in consequences affecting all members of the group in uniform ways. However, functional separation would not permit the direct use of individual records as the basis for individual actions. Alternative legislative proposals for implementing the concept of functional separation are reviewed in the chapter.

CHAPTER III

Major Administrative Data Files

This chapter describes the general properties of most of the major Federal administrative record files containing statistically useful information pertaining to individuals or businesses. The discussion is based largely on a survey of selected Federal agencies conducted by the SUAR Subcommittee. An attempt is made to lay the groundwork and indeed begin the discussion, continued in Chapter IV. of the statistical uses of administrative record systems.

Organizationally, the chapter is divided into four sections and two appendices. The first section indicates the scope of the administrative record files covered and describes the survey instrument used to obtain file documentation. In the second section there is a brief summary of the survey results. In the third section there is a brief description of the Social Security Administration's Continuous Work History Sample files. The CWHS files illustrate the process of extracting and merging information from basic administrative files to obtain files useful for

statistical analysis. In the final section there is a discussion of selected factors associated with the historical evolution of the statistical use of administrative files covered in the chapter. The survey questionnaire is reproduced in the first appendix, and a more detailed description of the CWHS program and data files is contained in the second appendix.

A. Scope of Study and Survey Conducted

1. 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 III.1. Information relating to individuals was sought from ten Federal agencies; some twenty-four administrative record files were involved in all.

Figure III.1 Major Administrative Record Files Surveyed 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	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	Individual Master Filer
Department 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	Summary Earnings Record
	Master Beneficiary Record
	Numerical Identification File (SS-3)
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
Bureau of Labor Statistics	Unemployment Insurance Address File
Department of Agriculture File	Producer Name and Address Master
	Economics, Statistics, and Cooperative Service List Sampling Frame
Department of Health and Human Services	Master Facility Inventory
Internal Revenue Service	Business Master File
	Exempt Organization Master File
Social Security Administration	Master Employer Name Directory
	Multi-Unit Code File

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 was also treated as "in scope" for comparison purposes. this time with business administrative record files.

2. Survey Conducted

In late 1978. the Subcommittee conducted a survey of the administrative files listed in Figure II.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.

A questionnaire was 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 inquiries about the files.

This survey collected data on both individual and business files by providing optional sections to 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.

B. 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 included in Tables II.1.1 and III.2.

1. 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 item included and the reporting units used:

a. Universe

In terms of coverage of individuals in the U.S. population. the decennial Census files are the most complete, followed by Social Security's Summary Earnings Files and the IRS Individual Master Fide. 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-for Federal government workers; and the Military Personnel Data Files-for present and former Armed Forces members.

b. Geographic information

Administrative files tend to have limited coded geographic information. Some contain a State code, but this was usually derived from the mailing address. The only exceptions appear to be SSA's Master Beneficiary Record file, and the related HCFA Health

Insurance Master 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 we 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 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 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 cost of this

straightforward approach makes it unlikely that it will be widely implemented on other files.

c. 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.

d. 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. 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.

2. Files pertaining Mainly to Businesses

The employer identification number is a major identifier on most of the administrative record files- including even the Census' Standard Statistical Establishment List. Some other similarities and differences in the files are:

a. Universe

The file with the largest coverage is the Master Employer Name Directory with about 27 million records' However, this file is not current and contains inactive businesses. The 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 Cooperative Service List Sampling Frame have extensive coverage of the farming sector.

b. Geographic information

As with the individual record systems, them is no subcounty geography data,present on any of the business files with the exception of the 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 headquarter's location can introduce a significant error into the data.

c. Economic data

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

d. Reporting Unit

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

C. Continuous Work History Sample Files

The survey results in the previous section indicate clearly that individual administrative record files usually do not contain the comprehensive population coverage and detailed identification of population characteristics desired for most

statistical analysis. The results also indicate, however, that it is often technically possible to overcome some of the limitations of single administrative files by linking several files and merging the information contained in these files. With files pertaining to individuals the SSN provides the principal basis for linkage and with business files the EIN is usually the basis for linkage. Both the problems and the potential benefits of file linkage we increased significantly when interagency linkages are considered (see, for example, the discussion of the Linked Administrative Statistical Sample in Chapter VI); but highly valuable statistical files can be developed through intra-agency linkages of administrative files in such large agencies as IRS and SSA. The Continuous Work History Sample program of SSA illustrates well the problems and potential of such intra-agency file linkages.

The CWHS program involves the construction of several statistical sample files from information contained in the SSA administrative files documented in Tables III.1 and III.2., The 1 percent 1937-to-date CWHS file, for example, involves primarily the extraction and merger of information from the Summary Earnings Record and Master Beneficiary Record files documented in Table III.

1. Annual and longitudinal employee-employer CWHS files are constructed largely by merging detailed earnings items which are input to the Summary Earnings Record File with industrial and geographic information obtained from the SSA employer files documented in Table III.2.

CWHS files do not contain occupational information for workers, nor do they contain the detailed socioeconomic characteristics available in census sample files. CWHS files do, however, contain information on worker sex, age, and race; and they can provide much greater longitudinal detail relating to the earnings history of workers than is available from any survey source. The CWHS program, moreover, has a considerable advantage over household surveys in obtaining employer information because of the possibility of direct links between employer and employee administrative files. The advantage of direct links between employer and employee information; however, is offset somewhat by quality problems associated with the geographic and industrial coding in SSA employer files (sec Chapter VII).

Because the CWHS program illustrates well both the potential and the problems associated with the statistical use of

administrative records. examples of CWHS applications and deficiencies are presented throughout the report. Some of the more detailed references to the CWHS program are included in: (1) the discussion in

Chapter V of the new joint IRS-SSA system of annual employer reporting (on Form W-2) of individual worker wages; (2) the discussion in Chapter VI of the development of the new Linked Administrative Statistical Sample program; and (3) the discussion in Chapter VII of technical problems encountered in the statistical use of administrative records. To permit the reader to better follow the references to the CWHS made throughout the report, a detailed description of the CWHS program and CWHS files is presented in the second appendix to this chapter.

D. The Evolution of Statistical Use of Administrative Records

Chapter IV contains a detailed discussion of statistical uses of administrative records from the perspective of selected Federal agencies that make extensive use of administrative records in their statistical and research programs. Chapters V and VI then follow with detailed case studies of selected projects and programs involving intensive statistical use of administrative records. To provide additional background for the chapters on uses, this section reviews some of the circumstances surrounding t he evolution of statistical uses of administrative record files covered in Tables III.1 and III.2.

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 (see Chapter IV). 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.

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 (NBER, 1958). After each decennial population census since then, there have been attempts to understand and quantify any error in the results by matching a small sample of census records to various administrative record sets such as IRS data (Schneider and Knott. 1973), Medicare data (U.S. Bureau of the Census 1973c), birth records (U.S. Bureau of the Census, 1963 and 1973a), death records (Kitagawa and Hauser, 1973), and employment records (U.S. Bureau of the Census, 1965). 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-see Chapter VI.)

With the extensive computerization of administrative files in the 1960's, the possibilities for expanded statistical uses 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 (Hansen, 1974). 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 (Kilss and Scheuren, 1979). 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). Adminis-trative 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 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 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 (Fay and Herriot, 1979). In this way. the definitional problems were minimized.

With the expanded interest in administrative records, them 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 (Word 1978). 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 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. (Chapter VII contains a more detailed discussion of the problems encountered in the BEA journey-to-work study.)

Currently, the Census Bureau uses IRS adjusted gross income and wages and salary 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 income size distribution estimates with IRS data has long been considered desirable. The inability to group IRS returns directly into families or households makes such updating difficult, but synthetic estimation procedures involving IRS data are being used in the development of family personal income size distribution estimates at BEA (see Chapter IV).

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 francs 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.

E. Appendix III.1

The Survey Questionnaire

Statistical Use Survey of Records Pertaining to

Individuals, Individual Firms, or Employers Maintain

and/or Mandated by the Federal Government

Survey for: Subcommittee on Statistical Uses of Administrative Records

Federal Committee on Statistis Methodology

Office of Federal Statistical Policy and Standards

Please complete the following questions as applicable. Since this survey covers individuals, householdsm and business organizations (firms and employers), not all of the questions may pertain to the data file you are answering the questions about. If you have any questions concerning the survey or concerning a particular question; or need additional copies of the survey form, please contact Ms. Maria Gonzales on (202) 673-7953.

(Please mark the appropriate category or categories

or supply the requested information)

1. What is the name of the file?

A) General name by which the file is usually called_____

B) Technical or official name if different from

the general name_____

2. What type of documentation exists for the file?

__ International Documentation

__ Not available to anyone outside the agency.

__ Available on request.

20

_ Outside Documentation

_ None currently prepared.

_ Available on request.

_ Not now available, but could be prepared upon request.

3. What type of documentation is available outside the agency?

_ Record Layout

_ File description--technical description

_ General file description without specific field description

_ No documentation available outside agency

4. What type of information is present on the file? The purpose of this question is to obtain a list of the kind of information present on the file which might have statistical uses. You may respond to the appropriate questions below or provide a separate listing of the information on the file. Is the reporting or filing unit an individual, household, business, or some other unit?

_ Individual (Answer 4A)

_ Household, Family, or Other Group of Individuals (Answer 4B)

_ Business or Employer (Answer 4C)

_ Other reporting unit (Answer 4D)

4A. What kind of information on individuals is present on the file?

	Please Circle Yes or No as Appropriate	
1) Person's name	Yes	No
2) Mailing address	Yes	No
3) Residence address	Yes	No

4) Has the address been assigned a geographic code? If yes, what level of geography are present? Yes No

State Yes No

County Yes No

Place Yes No

Other, please specify_____

5) Race--If yes, what are the categories? Yes No

6) Spanish or oher ethnic origin designation--If yes, what are the categories? _____ Yes No

7) Date of birth or age Yes No

8) Sex Yes No

9) Marital Status--If yes, what are the categories?_____ Yes No

10) Income--If yes, what are the types of income present?_____ Yes No

11) Person's family or household income--If yes, please specify type.

12) Social Security or Railroad Retirement Number — Yes No

13) Is the person's employer identified? Yes No

If yes, is the employer's Empoyer Identification Number present

14) Is the person's occupation identified? Yes No

15) Is the person's occupation identified? Yes No

16) Level of education or technical skill Yes No

17) Place of birth or foreign country of birth Yes No

18) Information on person's health or disability--If yes, please specify Yes No

_____

19) Other relevant statistical information --If yes, please specify_____ Yes No

4B. What kind of information on a household, family, or other group

of individuals is present on the file?

Please Circle Yes

or No as Appropriate

1) Person's name	Yes	No
2) Mailing address	Yes	No
3) Residence address	Yes	No
4) Has the address been assigned a geographic code? If yes, what level of geography are present?	Yes	No
State	Yes	No
County	Yes	No
Place	Yes	No
Other, please specify_____		
5) Household or family size	Yes	No
6) Each household or family member identified	Yes	No
7) Household or family income	Yes	No

The following questions apply to the household or familly head or primary applicant.

8) Date of birth or age Yes No

9) Sex Yes No

10) Race--If yes, what are the cate Yes No

gories? _____

11) Spanish or other ethnic origin des- Yes No

ignation--If yes, what are the

categories? _____

12) Social Security or Railroad Retire- Yes No

ment Number

4C. What kind of information on business organizations or employers

is present on this file?

	Company or Enterprise	Establishment	Employer Identification Number (EIN)	Other please specify in the Remark section
The file is organized by (please check	ß	ß	ß	ß

the correct):

1) Name	Yes	No	Yes	No	Yes	No	Yes	No
2) Address	Yes	No	Yes	No	Yes	No	Yes	No
3) Location code for establishment or other reporting unit	Yes	No	Yes	No	Yes	No	Yes	No

4C. What kind of information on business organizations or employers is present on this file? (Continued)

	Company or Enterprise	Establishment	Employer Identification Number (EIN)	Other please specify in the Remark

section

_____

4) Number of employees--	Yes	No	Yes	No	Yes	No	Yes	No
If yes, as of what date?_____								
5) Total payroll	Yes	No	Yes	No	Yes	No	Yes	No
Annually	Yes	No	Yes	No	Yes	No	Yes	No
Quarterly	Yes	No	Yes	No	Yes	No	Yes	No
6) Primary industry-- if yes what industry coding system is used? for example, 4 digit SIC, 2 digit SIC, etc. _____ _____ _____	Yes	No	Yes	No	Yes	No	Yes	No
7) Secondary industry	Yes	No	Yes	No	Yes	No	Yes	No
8) Gross sales or receipts	Yes	No	Yes	No	Yes	No	Yes	No
9) Product description	Yes	No	Yes	No	Yes	No	Yes	No
10) Amount and description of	Yes	No	Yes	No	Yes	No	Yes	No

capital base, total invest-

ment in plant and equip-

ment

11) What other items of statistical interest are available? Please list

in Remarks section below.

4D. What kind of information is available for the "other reporting unit?"

Please specigy the kind of information present on the file for the "other

reporting unit" in the space provided below.

5. What are the applications or forms which the data are derived? If

possible, include the OMB (or other) form number.

6. Briefly describe the process by which this information is obtained

from the individual or business(firm, employer) and procesed

to the data file being described.

7. What is the purpose of the file? If the purpose is to meet specific

legislative requirements, please include a citation for applicable

Federal law agency regulation, or agency requirement.

8. a) Is the file a computerized version of a "paper system?"

Yes No

b) What year was the file first created?_____

c) Has the file been expanded or has the data on the file

changed significanlty over its history? Yes No

If yes, please explain how.

9. How many individuals or businesses are represented on the file?

(An approximate number only.) _____

10. What are the restrictions on the use of file?

a) Legal Restrictions--

b) Administrative Restrictions--

c) Other Restrictions--

11. If either the SSN or EIN are present on the data file, what is their

purpose?

12. Is the file currently being used for statistical purposes?

Yes No

For example: Is the file used as a sampling frame for any surveys?

Are tabulations prepared from the file that are used for statistical

purposes?

Please briefly describe any statistical uses of the data file.

13. How often are data collected and updated for this file?

Collected Updated

_ One time only _ As needed

_ Annually _ Annually

- Quarterly _ Quarterly

_ Other, please specify _ Other, please specify

14. Please provide the name, address, and telephone number of a person who could answer questions concerning the data file (this persons need not be the same person who answers this survey).

Name: _____

Address: _____

_____

City and State:_____

_____

Zip Code: _____

Telephone Number: _____

15. Name and telephone number of person who completed this survey if different from above.

Name: _____

Telephone Number: _____

F. Appendix III.2

The CWHS Data System

The Continuous Work History Sample is a system of general multipurpose statistical data files designed primarily for socioeconomic research. The system consists of samples of records of individuals with employment covered by social security. Earnings, employment and benefit data for the individual along with personal characteristics and employer characteristics are maintained at varying degrees among five basic data files and two special files that are produced in the CWHS system.

This appendix describes: (1) the data sources for the CWHS system; (2) the procedures used to construct the administrative

data files underlying the system; (3) the procedures used to create statistical files from the records in the administrative files; (4) the sample design used for the system; and (5) the principal data elements in each of the five basic CWHS files. The discussion refers to data and procedures predating the start of annual wage reporting in 1979 (for calendar year 1978). A discussion of the new annual reporting system is presented in Chapter V. And Chapter VII contains considerable discussion of the limitations of CWHS data.

1. Data Sources

Data for the CWHS are obtained from records derived from reporting and informational forms and applications used in administering the retirement, survivors and disability programs of the Social Security Administration. The date of birth. sex and race of the person is obtained from the Application for a Social Security Number (Form SS-5). Geographic and industry information is obtained from the employer's Application for an Identification Number (Form SS-4) and other related forms that are used

periodically to update this information (Form OAA-100, OAA-103 and SSA-5019). Initially, employers are assigned geographical and industry classifications based on the location and nature of business information sup. plied on the Form SS-4. Information that is not satisfactorily reported on the SS-4 is obtained through the supplemental forms OAA-100 and OAA-103.

Employers who operate more than one place of business and have a total of 50 employees with at least six in a separate location are asked to use the Establishment Reporting Plan. Under this plan the employer gives SSA- a list showing the location. industrial activity and approximate number of employees of each establishment. On subsequent wage reports the employer groups his employees by establishment, identifying each group with a preassigned establishment number. The arrangement allows SSA to properly classify the employees according to geography and industry.

Data on earnings and employment are derived from various reporting forms submitted by employers and self-employed persons. Prior to 1978, with the advent of annual wage reporting, taxable wages of employees were reported quarterly by regular employers on Form 941, household employers on Form 942, and State and local

government employers on Form OAR-S3. Farm employers report annually on Form 943 and self-employed persons use Schedule SE of Form 1040 to report annually. (Refer to Chapter V for a discussion of the new annual reporting system).

Claims and benefits information is obtained from applications and forms that are completed in the process of filing for and determining entitlement to benefits.

2. Processing Procedures--Administrative Records

The demographic information (date of birth, sex and race) furnished by the applicant on the Form SS-5 is extracted after the social security number has been issued. This information is maintained on magnetic tape in a master file called the Summary Earnings Record (see Table III.1). This is the record in which the lifetime earnings and quarters of coverage of the individual is recorded for use in determining entitlement to benefits and calculating benefit amounts at the time a claim for benefits is made.

The information supplied by the employer on the Form SS-4,

relating to the location and nature of his business, is manually coded with geographic and industry codes. This information is key punched and maintained on magnetic tape in a master file of employers called the Employer Identification file (see Table III.2). Additionally, the information supplied on Form SSA-5019 by multi-unit employers using the Establishment Reporting Plan pertaining to the location and nature of business of each separate reporting unit, is also manually coded with geographic and industry codes and maintained in the EI file.

The earnings data that are reported by employers are received and processed at SSA in a variety of ways. Hand filled paper forms that meet certain criteria are optically scanned to produce a machine-readable record, while others are keypunched. Some employers, usually having a large number of employees, report directly on magnetic tape. The reports of self-employed persons are received directly from the Internal Revenue Service on magnetic tape. After all of the earnings data is in machine-readable form with appropriate identifying information, the tapes enter a computer balancing operation in which each page of each report is checked to see that the wage items balance to the page totals provided by the employer. Out

of balance items are investigated and corrective action taken. Balanced items are passed on to an operation where individual items are sorted in social security number sequence and then matched to the Summary Earnings Record on number and the first six letters of the surname. Earnings amounts are added to the summary records where complete matches occur. Unmatched records are rejected for further investigation and processing.

Prior to annual reporting, this processing occurred at regular intervals four times during the year. It generally takes about 9 months after the end of reference period to receive, process and update the summary earnings records with virtually all of the items for that period.

Claims for social security benefits are filed in local social

security district offices. Requests for earnings records and benefit computations are made by the district offices to SSA headquarters. After the earnings record is located, benefit computations are made and documentation of the claim is prepared and forwarded to the requesting office where the claim is developed and forwarded to program service centers for benefit authorization. Upon authorization of benefits, the program service center sends a notification of award to headquarters where a new beneficiary record is established in the Master Beneficiary Record file (see Table III.1). Changes to records in the beneficiary file are made through reports by the district office or program center. The Master Beneficiary Record file is used in the preparation of monthly social security benefit check records which are forwarded to the Treasury Department for payment.

3. Processing Procedures-Statistical Records

Once a year after the Summary Earnings Record has been updated with virtually all of the prior year's earnings, a 1 percent sample (based on specified digits of the social security

number) is extracted. This file becomes the foundation for producing the 1 percent 1937-to-date CWHS. It is used along with the prior year's CWHS, a 1 percent sample extracted from the Master Beneficiary Record file, and miscellaneous correction files to generate the required data elements for the current year's 1 percent CWHS.

At the same time that earnings data for the current processing period are posted to the Summary Earnings Record, the 1 percent sample of earnings items records are written off separately on magnetic tape. The items are accumulated until all four quarters of the year have been processed. They are then summarized into one record for each employee-employer-establishment combination with quarterly earnings amounts maintained separately. The resulting records are matched to the Employer Identification file and geographic and industry codes are inserted. They are then resummarized to an employee-employer level. Cases having employment with more than one establishment of the same employer are assigned to the unit having the most activity in terms of quarters of employment. A match is the n made to a special extract from the 1 percent sample 1937-to-date CWHS containing date of

birth, sex and race codes. These personal characteristics are inserted into the record to form the final 1 percent Sample Annual Employee-Employer file.

Another file of the earnings items that are posted to the Summary Earnings Record, previously referred to, is written off separately for another type of processing. This is a 0.1 percent sample and is a subset of the 1 percent sample. These records are accumulated over the same time period as the 1 percent sample records and are processed along with the prior year's 0.1 percent basic file and a special 0.1 percent write off of certain data items from the current year's 1 percent CWHS file to create the current year's 0.1 percent 1937-to-date CWHS.

Information for self-employed persons. coming from the Schedule SE of the Form 1040, is submitted to SSA from IRS directly on magnetic tape. After initial processing of these records in order to properly credit and post earnings to the Summary Earnings Record, the 1 percent sample records in this file are written off for statistical processing. In subsequent computer operations IRS industry codes that are in the original record are converted to SSA industry codes and addresses are converted to geographic codes through a special coding file that utilizes Zip code and place

names. Correspondence is generated for cases with missing and/or incomplete information asking for the required data. The final resulting file from these operations is the 1 percent Sample Annual Self Employed file.

In addition to the regular statistical processing described above, in recent years special processing has been done to generate two additional files; the First Quarter Employee-Employer-Establishment files for the 1 percent sample and a special 10 percent Sample First Quarter Employee-Employer-Establishment file. Processing for these files is similar to processing for the Annual Employee-Employer files except that it is done after all first quarter receipts have been received and posted to the summary earnings record. Record contents are virtually the same as the annual except that only first quarter data are included. The 1 percent first quarter files have been prepared for the years 1970-76, while the 10 percent first quarter files have been produced for the years 1971, 1973, and 1975.

4. Sample Design

The population from which the CWHS is selected consists of the one billion possible nine-digit social security

numbers. These numbers have the following digital arrangement:

Area in which number assigned (three digits)	Group number (two digits)	Serial number (four digits)
XXX	XX	XXXX

In the issuance of social security numbers, each State is assigned one or more area numbers with the exception of a special block of numbers assigned prior to August 1963 to persons covered under the

Railroad Retirement Act. Each State number, in combination with a given group number defines a stratum. The population assigned social security numbers is thus stratified geographically (by place of application for social security number) and chronologically (by the process of assigning these numbers). Each number is an element of a given stratum, and the population represented by the possible one billion elements constitutes the sampling frame.

The CWHS is a longitudinal sample of persons with covered employment. The sample consists of all persons who have social security numbers with specified digits in certain of the serial-number positions and who have covered employment during any defined reference period. The digital selection pattern remains constant. The employment and earnings histories for persons in the sample are available from 1957 forward, with limited additional earnings data going back to 1937.

The 1 percent CWHS may be described as a stratified cluster probability sample of all possible social security numbers. A stratum consists of all social security numbers with the same area-group number. In a stratum for which all numbers have been issued, the 1 percent sample consists of 100 of the 9,999 social security

numbers issued. (Numbers ending in 0000 are not assigned.)

The clustering within a stratum arises from the particular digital selection procedure used, in combination with past methods of assigning social security numbers. Because of the clustering, sampling errors of estimates from the 1 percent CWHS are slightly larger than those that would result from a stratified random sample of the same size.

The present design of the 1 percent sample evolved from earlier sample designs--an initial 20 percent sample and a later 4 percent sample. All past designs have used the same stratification modes as are used in the present design.

The 10 percent CWHS is a stratified systematic sample. The strata are the same as those used for the 1 percent sample, and the digital selection procedure within strata is such that them is no clustering effect. Therefore, sampling errors of estimates from the 10 percent CWHS are presumed to be about the same as or slightly smaller than those that would result from a simple random sample of the same size.

5. Data Files

A brief description of the files produced in the CWHS system is shown below, including a listing of the major data elements. These files had been made available on a cost reimbursable basis with precautions taken to preserve the confidentiality of information relating to specific individuals or reporting units. These precautions included limiting the data elements to those needed by the researcher for the purposes stated and transformation of identifying numbers to unique case numbers which still permit linking of common records among various files. Additionally, a conditions-of-release agreement was signed by the requestor. At present, however, SSA is not releasing CWHS files to the public pending legal clarification of restrictions on release imposed by the Tax Reform Act of 1976.

a. One percent sample annual Employee-Employer (Ee-Er) File

A 1 percent sample of social security numbers for which wage and salary employment was reported in the reference year. There is one record for each employee-employer combination. Basic data

elements: (1) personal characteristics--year of birth, sex, race; (2) wages-annual taxable, quarterly taxable, and total estimated wages; (3) employer-State and county, industry. coverage group (farm, household, Federal civilian, etc.); (4) insurance status; (5) benefit status.

b. One percent sample annual Self-Employed (SE) file

A 1 percent sample of social security numbers for which self-employment earnings subject to social security coverage were reported in the reference year. Basic data elements: (1) personal characteristics-year of birth. sex, race; (2) self-employment--taxable income, net comings, State and county, industry; (3) taxable earnings (including wages, if any); (4) type of work-farm or nonfarm self-employment (and wage indication. if any); (5) insurance status; (6) benefit status.

c. One percent sample Longitudinal Employee-Employer Data (LEED) file

Assembled from the 1 percent sample annual Ee-Er records which art prepared yearly. In the annual files. one record is created for each employee-employer combination during the year. In the longitudinal file, the original records from the various annual files have been skeletonized, resequenced, and merged so that all records associated with an employee over the time span of the file appear

together. Basic data elements are the same as in the 1 percent sample Ee-Er.

d. One percent 1937 to date CWHS file,

A 1 percent sample of social security numbers issued

through cut-off date of file reflecting entire work experience in covered employment. Basic data elements: (1) personal characteristics- year of birth, sex, race; (2) employment-number and pattern of years employed, first and last years employed, pattern of quarters employed (last 2 years), number of quarters of coverage 1937 to date, pattern of quarters of coverage 1957 to date; (3) type of work-farm or nonfarm, wage or self-employment; (4) taxable earnings each year 1951 to date; (5) self-employment-- taxable income each year 1951 to year prior to current year, net earnings, for year prior to current year; (6) insurance status; (7) benefit status.

e. One-tenth of 1 percent 1937 to date CWHS file

A 0.1 percent sample of social security numbers issued through cutoff date of file reflecting entire work experience in covered employment. Basic data elements are generally the same as for the 1 percent CWHS except for more detailed earnings information, e.g., taxable wages each year 1937 to date, taxable farm wages each yew 1955 to date, quarterly wages each quarter of each year 1951 to date, net earnings from self-employment each year 1956 to date.

In addition to the files described above, two others have been created at the request of the Bureau of the Census and the Bureau of Economic Analysis-the 1 percent sample and 10 percent sample First Quarter Employee-Employer-Establishment file. Microdata has been made available from the 1 percent sample first quarter file; however, only summary files and tabulations from the 10 percent sample are available.

CHAPTER IV

Major Statistical Uses of Administrative Records

Most of this chapter is devoted to review of statistical uses

of administrative records in five selected Federal agencies. These agencies include: (1) the Internal Revenue Service and (2) the Social Security Administration, which represent two of the largest primary collectors of administrative data pertaining to individuals and businesses. (3) the Bureau of Economic Analysis, which uses administrative record data extensively in making estimates for the national economic accounts and related statistical series; (4) the Bureau of the Census, which uses a wide variety of administrative records in developing sampling frames and evaluating survey data as well as directly in estimating statistical-series; and (5) the Small Business Administration, which is in the process of using data from a variety of administrative sources in the development of a general-purpose small business data base for use in research and policy analysis. Although them is no review of administrative record use of the Bureau of Labor Statistics in this chapter, Chapter V contains a major case study involving BLS use of administrative records from the Unemployment insurance payroll tax system.

The discussion of uses in this chapter is not intended to be comprehensive. Brief overviews of uses by agency are supplemented

by a few more detailed discussions of uses in specific programs. The more detailed discussions involve primarily Census Bureau programs in the area of economic statistics. A number of Census Bureau uses of administrative records in population statistics programs are covered in some detail in other chapters (especially Chapter VI).The overviews of IRS and SSA programs are brief, but examples of uses of IRS and SSA administrative records appear repeatedly in other chapters. The narrative discussion of BEA uses is brief, in part because many of the uses of administrative records in economic accounts can be conveniently summarized in tabular form. The SBA discussion involves a new program still under development and is intended primarily to illustrate problems facing the development effort.

Chapter III has already provided some selected examples illustrating the historical development of statistical use of administrative records. As with the examples cited in Chapter III, most of the examples considered in this and subsequent chapters involve direct or indirect use of primary administrative files such as those documented in Chapter III. The distinction between administrative and statistical data files, however, has not always

been made clear. Therefore, to provide some additional perspective on the process of making statistical uses of administrative records, the first section of this chapter discusses some of the general considerations involved in defining administrative record files and in creating and using statistical files derived from administrative records. Following the first section, the remaining five sections of the chapter discuss uses of administrative-based statistical data on an agency-by-agency basis. An appendix contains selected tabular materials relating to the agency discussions.

A. Defining Administrative Record Files and Using Them statistically

In statistical uses of records pertaining to persons or businesses, the interest is generally in studying the characteristics and attributes of groups of individual entities as opposed to identifying specific entities and taking actions based on their individual characteristics as in administrative uses. Indeed, in censuses and surveys involving direct collection of information for statistical use, it is usually felt to be important

to provide assurances to participating respondents that information they supply will not be used as a direct basis for administrative actions against (or for) them specifically. Therefore, in this report statistical (as opposed to administrative) record files will generally be considered to be files which are not made available for taking administrative action with respect to individual legal entities (persons or businesses); i.e., files which are not used to determine an individual reporting entity's legal obligations or benefit entitlements.

Given the distinction between statistical and administrative files just suggested, it should be acceptable to create statistical files from administrative files, but not vice versa. This concept of "functional separation" of records is being considered in proposed legislation (see Chapter VIII), and is applied in SSA's Notice of Proposed Rulemaking to revise its Regulation No. 1, but is not yet well established in either the regulations or the procedural policies followed in many Federal agencies. The result is

considerable variation and confusion in the extent to which administrative records can be made available for statistical uses. Problems related to limitations and confusion surrounding access to administrative records for statistical use will be discussed in connection with examples covered throughout the report; and the legal aspects of the access issue will be reviewed in detail in Chapter VIII. The remainder of this section provides a brief, but somewhat more general overview of considerations associated with using administrative records for statistical purposes.

The primary distinction between administrative records and statistical records is the ultimate use to which they are intended to be put. This usually means a parallel distinction in the degree to which the statistician is in control of the design and collection of the records. Survey records and their collection procedures are designed, documented and controlled to yield the

desired statistical characteristics. When administrative records are used statistically, the statistician must locate existing records and determine their conceptual suitability for the intended use. And the statistician must also devise methods for overcoming technical problems frequently encountered in making new uses of existing records.

As noted in Chapter III, most statistical uses of administrative records have developed on an ad hoc basis. With the exception of uses by the collecting agency to generate statistics needed for program administration, there are few examples of administrative record systems that have been designed with statistical uses in mind. In most instances the statistician, faced with the problem of generating statistics for a particular policy analysis, fund distribution, or program evaluation purpose, has approached an administrative record system from the standpoint of what is available for the current application. in some instances these ad hoc uses have become regularized and institutionalized, but only rarely have statisticians specified changes in the design or procedures of an administrative record system necessary to yield more reliable statistics. This is true even when the statistical

analysis provides essential feedback for the operation of the administrative system.

Statistical uses of the various administrative record sets have generally been uncoordinated. A body of uses and users have developed independently for each record set. For this reason, and because the records are collected by different agencies with differing legislation governing their collection and use, there is very little standardization of the accessibility, documentation, format, and quality of information available from the various record systems.

Statistical uses of administrative records, moreover, are often met with some resistance from the operating personnel of the collecting agency. This is partially due to diffusion of responsibilities. Organizations which have responsibilities for assembling statistics are usually no the same as those which have responsibilities for maintaining administrative records and consequently producing and using agencies have differing priorities. Even the statistical units of administrative agencies are primarily responsible for meeting the statistical needs of that program and only secondarily for meeting the statistical needs of other Federal agencies, State and local governments, and other

public and private concerns. Statistical uses are often viewed by administrative personnel as an annoying addition to their already overburdened work schedules.

Other reasons for this resistance are related to confidentiality restrictions and the massive nature of the record sets. Many of the record sets are collected with either formal or informal assurances of confidentiality to the participating entities. Administrative personnel are therefore either unable or reluctant to make the records available for statistical use. Many of the record sets are so large, amounting to many millions of records, that even a seemingly minor change in the information to be collected or the collection and processing procedures could have far reaching cost and timing repercussions.

B. Internal Revenue Service

The Internal Revenue Service, in its role as tax collector, acquires millions of records from nearly all units of the economy: individuals, proprietorships, corporations, and nonprofit

institutions. These records are collected for tax-administration rather than statistical purposes. They are, however, used to generate a wide variety of statistics. The Statistics Division of the IRS has responsibility for assembling statistics from tax records. These statistics are used for program planning and many are also published for general use.

The program planning uses range from analyses of simple operating statistics, such as the number of returns processed and taxes paid, to analyses of alternative tax policies, including the assessment of revenues that would be raised under alternative policies and the impact of those policies on the economy.

The publications for general use include the Statistics of Income reports (annual) based on individual, corporate and other business tax returns; occasional reports based on information obtained from fiduciary, estate, foreign and other tax returns and schedules; and first-time reports (in preparation) on finances of tax-exempt organizations and pensions plans. Supplemental reports are prepared biannually which classify information from individual returns by SMSA and by county. These reports are used to provide basic information for tax studies by Congress and

its committees, for administrative use by the Secretary of the Treasury and the Commissioner of Internal Revenue, and by other Federal agencies, as for example, in BEA's construction of national and regional economic accounts. They are also used for general economic research in the areas of income and wealth.

Many of the IRS statistical series are produced from samples of tax returns. The sample files, devoid of identifying information, are made available to bona fide researchers on a cost reimbursable basis. The appendix includes a description of the major administrative record files maintained by IRS. as well as a list of Statistics of Income publications.

The extensive statistical use of IRS records is indicated not only by the diversity of IRS publications and internal programs,

but also by the prominent role of IRS records and tabulations in the uses to be discussed later in this chapter for the Bureau of Economic Analysis, the Census Bureau, and the Small Business Administration. In addition, IRS data play prominent roles in many of the case studies examined in Chapters V and VI.

C. Social Security Administration

The Continuous Work History Sample statistical program of SSA has already been discussed in Chapter III. But the CWHS program emphasizes work-related data from its payroll tax program much more than data connected directly with SSA disbursement of benefits under its various programs. In addition to regular Old Age, Survivors, and Disability Insurance benefit programs, SSA also administers the Supplemental Security Income program for the needy, aged, blind, and disabled and the Aid to Families with Dependent Children program which provides financial assistance to certain qualified needy children; and until a reorganization within the Department of Health, Education, and Welfare in March 1977, administered the health insurance program under Medicare. The Medicare

program is now administered by the Health Care Financing Administration, but SSA continues to provide selected data processing services for HCFA. And SSA is also continuing to administer the distribution of certain black lung benefits to coal miners and their families. In this case SSA responsibility covers some new claims as well as those claims that were filed before the basic black lung program was transferred to the Department of Labor.

In administering these varied and complex programs, a great many records are maintained from which statistics are regularly generated. These statistics relate to general and specific aspects of the various SSA programs, dealing with number of claims, number and amount of benefit payments, post entitlement actions, administrative costs, etc.

Throughout the development of the social security system, research has been important to policy formulation and program administration. The Office of Research and Statistics is the chief research resource of SSA and has the responsibility for all program statistics and for analyses required by the Administration and by Congress. In carrying out its mission, ORS disseminates a large

volume of statistics in the monthly Social Security Bulletin and its Annual Statistical Supplement as well as in other reports, papers, and statistical releases. The appendix (section G.2) gives an illustration of the great variety of statistics that are produced by ORS. The tables listed there were taken from the table of contents of the 1976 Annual Statistical Supplement to the Social Security Bulletin.

D. Bureau of Economic Analysis

BEA relies heavily on administrative records in the preparation of national economic accounts and related measures. BEA's estimates of current economic activity are based, with few exceptions, on analysis of primary data obtained from other agencies. This use of available materials is economical because it does not require extensive primary data collection activities. It has the further advantage of not adding to the reporting burdens of businesses and individuals. The process does, however, place a burden on analysts in terms of adapting data designed for other

uses, remaining alert to changes. in source data-, and researching potential new data sources. In this dual role as an intensive user and producer of government statistics, BEA accumulates more experience than most other agencies with the systematic use of a wide variety of administrative records. The lack of consistent definitions and procedures, uncoordinated formats and presentation techniques, and inadequate timing are familiar to the BEA analyst who must be aware of and make adjustments for deficiencies in primary data.

The list of administrative record tabulations which are used directly to estimate components of the national income and product accounts, the national input-output tables, and the international accounts is extensive and includes various types of tax records, regulatory records, financial records of the Federal Reserve System and Federal Deposit Insurance Corporation, custom reports, and budget documents. Tables IV.1, IV.2, and IV.3 list the components of the NIPA, input-output accounts and international accounts which are based on administrative records. The tables also indicate the source of the records used. In addition, Table IV.4 lists components of the NIPA that are based on data from current surveys

for

which the sampling frames have been developed from administrative record sources. (The development of such sampling frames is discussed in the Census Bureau section).

BEA's estimates of State and local area personal income involve the use of many of the same administrative record sets indicated for components of national personal income in Table IV. 1. In fact, since most current statistical surveys have sample sizes that we too small to provide reliable State and local data, administrative records play a relatively more important role in State and local than in national personal income estimates. Tax records and budget documents are the most important sources. The Unemployment Insurance payroll tax program (see the case study in Chapter V) is the principal source of wage and salary data, IRS tax

returns are the principal basis for estimating most components of property income and nonfarm proprietors' income; and government disbursement and related records are the basis for estimating the bulk of transfer payments to individuals.

For most of its work, BEA uses tabulations of records maintained by other agencies rather than using microdata files directly. In a program to develop estimates of family personal income size distribution, however, BEA is working cooperatively with SSA in the use of statistical matching techniques to merge information from administrative-based microdata files with Current Population Survey records. The administrative data include SSA's summary earnings and benefit records and IRS records from the Individual Master File, Statistics of Income File, and Taxpayer Compliance Measurement Program File. An additional administrative-based microdata file used extensively by BEA, particularly in regional analysis, is SSA's employee-employer Continuous Work History Sample (see Chapter III). In each of these microdata files used at BEA. individual identifiers have been removed or scrambled" to protect confidentiality. Even so, BEA access to several key flies including the CWHS and TCMP files has been at least

temporarily halted by the Tax Reform Act of 1976.

Table IV.1 National Income and Product Account

Components Based on Administrative Records

NIPA Component	Administrative Record
Personal consumption expenditures:	
Tobacco and alcohol	Tax records of Bureau of Alcohol, Tobacco and Firearms
Medical and legal services	Business income tax returns
Brokerage charges	Regulatory reports of the Securities and Exchange Commission

Table IV.1 National Income and Product Account

Component Based on Administrative Records -- Continued

NIPA Component	Administrative Record
Bank service charges	Regulaory reports of the Comptroller of the Currency, Board of Governors of the Federal Reserve System and the Federal Deposit Insurance Corporation
Consumer share of new motor vehicles	State government motor vehicle-registration forms
Air transportation	Regulatory report of the Civil Aeronoautics Board
Other intercity transportation	Regularoty reports of the Interstate Commerce Commission
Change in business inventories:	

Book value of inventories of nonfarm industires other than manufacturing and trade	Business income tax returns
Net exports:	
Merchandise trade	Customs reports
Federal Government purchases of goods and service	Budget documents
Wages and salaries:	
Nonfarm	Empoloyer payroll tax returns
Federal government	Budget documents
Employer contributions to social insurance	Employer payroll tax returns
Other labor income:	
Pension plan contributions	Business income tax returns
Nonfarm proprietor's income	" "
Corporate profits	" "
Corporate profit taxes	" "
Dividends	" "
Capital consumption allowances	" "
Business transfer payments	" "
Net interest	Business income tax returns

	and regulatory reports
	of the FRB, FDIC,
	CofC, and Federal Savings
	and Loan Insurance
	Corporation
Indirect business taxes and subsidies	Various tax records
Transfer payments	Various budget documents

Table IV.2 Input-Output Account Industry Estimates Board on Administrative Records

1-0 Industry	Administrative Record
Agriculture, forestry, fisheries	
Receipts for use of national forest and forest services	Reports of the US Federal Service
Aerial application services	Reports of the FAA
Mining:	

Rental and royalty receipts	IRS, Statistics of.....

Table IV.2 Input-Output Account Industry Estimates Board on Administrative Records--Continued

1-0 Industry	Administrative Record
Constrution:	
Installed cost of construction	Regulatory reports of the ICC, FPC, FCC

Manufacturing:	
Addition of excise tax	Administrative reports of the Treasury
Addition of rents and royalties	IRS, Statistics of Income
Small firm coverage in economic census	Administrative records (Census)
Addition of competive imports	Customs data (Census)
Transportation:	
Operating revenues and expenses of:	
Regulated components of railroads, trucking, water and petroleum pipelines	Regulatory reports of ICC
Regulated air	CAB
Unregulated components	CAB, USDA, FAA, Corps of Engineers
Utilities:	
Operating revenues and	

expenses of regulated companies	Regulatory reports of FCC, FPC ETA, REA
Water and Sanitary Services	IRS, Statistics of Income
Wholesale and retail trade:	
Gross margins on sales	IRS, Statistics of Income
Sales and excise taxes and duties:	
Federal	Treasury reports
State and local	State and local administrative reports (Census)
Finance, insurance, and real estate:	
Banking and finance	FRB, FDIC, IRS Statistics of Income Administrative reports of Federally chartered banks and lending agencies
Insurance agents and brokers	IRS, Statistics of Income
Rents paid by business	IRS, Statistics of Income
Royalty receipts by business and persons	IRS, individual income tax returns
Rent and royalty receipts and	

payments by governments	Budget documents
Commissions for management	
and transfer of property	IRS, Statistics of Income
Other services:	
Activities outside the scope of	
economic censuses:	
Accounting, auditing, and	
other professional	
services	IRS, Statistics of Income
Medical services	IRS, Statistics of Income
Education service	
expenses	Office of Education
Government enterprises:	
Federal enterprises	U.S. Budget, Treasury Department and agency reports
State and local enterprises	State and local budget documents (Census)

Table IV.3 Balance of Payment Account Components

Based on Administrative Records

Balance of Payments Component	Administrative Records
Merchandise exports and imports	Customs-Census reports
Transportation	Customs-Census reports
U.S. Government miscellaneous services	U.S. Post Office Department; Department of Justice
Travel	Immigration and Naturalization Service; Department of Transportation; Civil Aeronautics Board; State Department; Bank of Mexico; Statistics Canada; Federal Reserve Board
Official reserve assets	U.S. Treasury
Claims and liabilities reported by U.S. Banks	U.S. Treasury; Federal

	Reserved System
Claims and liabilities on unaffiliated foreignes reported by U.S. non-banking concerns	U.S. Treasury; Federal Reserve System
U.S. Securities and foreign securities	U.S. Treasury; Federal Reserve System

Table IV.4 National Income and Product Account

Components Based on Current Surveys

Using Administrative-Record Based Sampling Frames

NIPA Components	Administrative Records
Personal consumption expenditures:	
Goods, less motor vehicles	Monthly Retail Trade Survey (Census)
Personal and professional	

services	Monthly Selected Services Survey (Census)
Producer's durable equipment	Annual Survey of Manufactures (Census): Monthly Manufacturers Shipments Survey (Census) Quarterly Plant and Equipment Expenditures Survey (BEA)
Structures	Construction put-in-place (Census)
Change in business inventories, manufacturing and trade	Monthly surveys (Census)
Wages and salaries	Monthly Establishment Survey (BLS)
Corporate profits	Quarterly Financial Report (FTC)

E. Census Bureau

The Bureau of the Census is the largest primary data collection agency in the Nation. It conducts the decennial censuses of population and housing, economic censuses, agricultural censuses, censuses of governments, special censuses and numerous sample surveys. In addition to these vast data collection activities, the Bureau is also a major user of administrative records. It uses them directly to tabulate time-series information and indirectly in a variety of ways including: design and evaluation of censuses and surveys; identification of sampling universe; estimates for non-surveyed portions of the universe; and imputations for missing cells.

The distinctions between administrative and statistical records become particularly blurred with the Census Bureau applications because so many of the records which we generally consider as statistical are derived from censuses or surveys which utilize administrative records in many important ways. Even the

decennial censuses have in the past, utilized administrative records in design and evaluation phases.

Chapter III has already noted a major Census Bureau administrative records program for developing intercensal population and per capita income estimates for use in distributing General Revenue Sharing funds to State and local areas. Chapter III also mentioned the importance of administrative records in evaluation programs for the decennial censuses. And Chapter VI contains three detailed case studies illustrating administrative record use in evaluation and improvement projects for the 1980 Census and in development plans for the proposed Survey of Income and Program Participation household survey. The examples of administrative record uses cited in the remainder of this section will be drawn primarily from areas of Census Bureau responsibility for developing business and economic statistics.

1. Economic Censuses

Under Title 13 of the United States Code, the Bureau of the Census is required to conduct a group of economic censuses at five-

year intervals in the years ending in "2" and "7", the latest one covering 1977. This group includes the Census of Manufactures (initiated in the year 1810), Mineral Industries (1840), Retail and Wholesale Trade and Construction Industries (1929), Selected Service industries(1933), Public Warehouses(1935), Transportation (1963), and beginning in 1977 the remaining Service Industries (Medical, Educational and Non-Profit Areas).

In order to minimize the cost of the censuses and relieve the business community of reporting burden, the Census Bureau makes extensive use, under strict confidentiality restrictions, of selected information derived from tax records. These records form an integral part of the preparatory and collection phases of the economic censuses. The universe of business firms is based on selected information extracted from tax records for a tax year period encompassing the census year. This information. received on computer tape includes (1) firm name and address; (2) identification number, (3) legal form of organization. (4) business activity code; (5) number of employees; and (6) payroll by quarter.

For the 1977 economic censuses, the above basic information was integrated with the Standard Statistical Establishment List

(see Chapter V) and other sources. This process provided an almost complete list of approximately 12,000,000 business firms engaged in economic activity in the United States (including social and professional services) classified by kind of business and approximate size with employers and nonemployers separately identified. For this universe. the following subgroups were identified:

1. Those 5,200,000 businesses that could be excused from filing any questionnaire because their kind of business as determined from tax records was not in scope of the economic censuses;
2. Those 3,800,000 in-scope small businesses that could be excused from filing any questionnaire since limited data (receipts, payroll) extracted from tax records could be used to develop equivalent census-type data;
3. Those 3,000,000 larger businesses that were engaged in activities in-scope of the Censuses. Direct reporting was required for these firms in order to obtain all the information needed for the census results.

Therefore data for approximately 56% of the total business establishments covered in the economic censuses are extracted from administrative records. Data for companies that were not canvassed are obtained from the following additional items of information extracted from tax records:

1. Employment
2. Payroll
3. Sales or receipts
4. Physical location (not available if left blank on tax forms)
5. Business status at end-of-year
6. Number of months in business

The cost of obtaining these extracts of tax records was less than $2 million out of the total economic census budget. The equivalent cost to the Government of obtaining census reports from the excused group of about 8.500,000 businesses would have been at least 10 times

that amount given the availability of a complete mailing list of the excused businesses.

The quality of statistics produced by this meshing of tax records with reports to the Census Bureau would likely result in more complete coverage than that obtained by full field enumeration or combinations of field and mail enumeration techniques. For example, the field-enumerated 1948 Census of Business undercounted the number of standard retail establishments by at least 150,000. The undercount of nonstore business (e.g., mail order. house-to-house, vending machine, and service businesses) was also substantial but could not be determined using standard post enumeration surveys. In fact, the latter group in many cases can only be identified from tax records. In addition to identifying the universe, data from IRS tax records are also used for companies

which fail to report and for editing the reported data provided by the respondent. (See Chapter VII for a discussion of quality problems with administrative-based statistics.)

2. Census of Agriculture

The Census of Agriculture, started in 1840 and taken at 5-year intervals beginning in the 1920's, is the only source of statistics on agriculture that are comparable, county by county, on a nation-wide basis for farms classified by size, tenure, type of organization, market value of farm Products sold, and type of farm enterprise. The census data are widely used by Federal, State and local governments in a variety of ways in the administration of various farm programs, as benchmarks for the current crop NW livestock estimates issued by the Department of Agriculture, and in the preparation of overall measures of the economy such as the input-output ut tables for the national economic accounts.

Prior to the 1969 census, data collection was by personal interview. Information copies were distributed by mail to all households on rural routes and to post office boxes in rural

communities in the effort to locate all farm operators and have them complete the report prior to its pickup by the enumerator. Correlated with the burgeoning increase in the size of farms, there has been continuing rise in the number of farmers who do not live on the farm they operate--that is, a growing number of operators for whom door-to-door enumeration is not a practical possibility. Furthermore, the availability of capable people willing to accept short-term employment as census enumerators has steadily declined, making it more and more difficult to recruit an acceptable field staff in all areas. Fortunately, the availability of farm-related mailing lists from administrative records had increased correspondingly and this factor was instrumental in redesign of the dam collection procedures.

In planning for the 1969 Census of Agriculture, it became evident that the method of data collection should be changed from personal interview to a mail enumeration procedure based on administrative records. The size measure contained in the administrative tax records was the controlling factor that enabled the Bureau to send abbreviated report forms to small farmers and thereby reduce the reporting burden for nearly one-half of the nation's farm operators. This resulted in an obvious reduction in

costs for collecting and processing the census data. Subsequent censuses, including the 1978 Census of Agriculture, which is underway, have benefitted from the experiences and results obtained from the 1969 undertaking where under-enumeration of small farms was a severe problem.

3. Survey of Minority-Owned Businesses (SMOBE)

In 1969, SMOBE was conducted as a special project and funded by various government agencies to determine the extent of business ownership by minorities. Beginning in 1972, SMOBE became a part of the economic censuses that are required by law every five years. SMOBE is issued in a four part series covering businesses owned by Blacks, persons of Spanish Origin, Asian Americans, American Indians and Other Minorities. Data published cover number of firms, gross receipts, and number of paid employees. Tax records are used extensively in developing the statistics. For example, minority ownership is measured for the segments of the business population using HRS corporation, partnership and sole proprietorship tax forms and Social Security

Administration race codes to identify businesses for "Whites", "Negroes" and "Other Minorities." A mail survey is required to determine businesses owned by persons of Spanish Origin and the specific minority groups included in the "Other" minority category. However, the mail survey is minimal compared to the effort and costs that would be involved if tax records were not available. (See Chapter VII for a further note on limitations of SSA race codes.)

4. Current Economic Indicators

In addition to the quinquennial economic censuses and the 5-year census of agriculture, the Census Bureau conducts a broad series of weekly, monthly, quarterly, and annual sample surveys in the industrial, distributive trades and service areas. Some of these surveys have been in existence for several decades and have been converted from a design based primarily on use of area samples.i.e., an enumerator canvass of businesses located in a sample of land area segments--to a mail canvass of sample of businesses selected from the comprehensive tax file of firms

classified by size and industry.

The samples used to collect information concerning the distributive and service mules are primarily drawn from a

list of employer firms obtained from administrative tax records and updated through reconciliation to the economic census results. The volatility of changes in the business universe, however, requires that the sampling be updated often, if possible every quarter, to include new business establishments and to delete those no longer in operation. This updating process is based on information received from IRS on additions to and deletions from its list of active businesses. The total list of businesses obtained from IRS source&serves; as a control to assure that the data compiled in fact fully cover the sectors surveyed.

In the current industrial statistics program, similar updating procedures from administrative records we followed but on a less frequent basis. This includes the annual survey of manufactures, the monthly survey of manufacturers shipments, inventories, and orders and the more than 100 other current industrial reports relate to specific commodity areas such as fats and oils, paper and paperboard, and steel. The availability of updated complete tax files has made it possible for the Bureau to undertake on very short notice special surveys designed to meet policy-makers' needs. Recently, for example, the Bureau undertook, at the request of the Federal Reserve, a survey of industrial capacity to improve the statistics relating to current business conditions. Surveys involving energy-related industries have also recently been instituted. In general, the availability of lists of businesses classified by industrial category provides the Bureau with great flexibility in meeting new or changed objectives.

5. The Standard Statistical Establishment List Program

The SSEL program is discussed in detail in Chapter V; but it should be noted here that the SSEL provides an important mechanism

for coordinating most of the economic censuses and surveys discussed above. In addition, County Business Pattern publications of employment and payroll data for State and local areas are now based directly on the SSEL.

F. Small Business Administration

Federal economic and business statistics have generally not been well designed for the analysis of small business. Many agencies do not prepare tabulations by size of business and them have been no standard guidelines for preparing size class data so that data available by size frequently cannot be readily compared or integrated across agency sources. Size class data, moreover, are often not available for comparable reporting units or on the basis of comparable size indicators. IRS corporate tax return data, for example, are available for tax paying units which differ from the establishment concept used in the preparation of most Census Bureau business data. Moreover, Census size class statistics usually do not distinguish between establishments that

are separate business entities and establishments that are a part of larger multi-unit companies, and most Census size class data use employment as the indicator of establishment size, whereas IRS business income tax returns collect no employment data and are traditionally tabulated by size using such alternative indicators as level of assets or business reports

To address the problem of inadequate data relating to small business, an interagency committee has recently been formed with a mandate from the President to establish a small business data base. SBA has been charged with the principal responsibility for assembling the data base. And because of the high paperwork costs to small businesses of detailed Federal business reporting requirements, emphasis in developing the new data base will be placed on utilizing existing primary data sources and particularly on more efficient statistical use of administrative records. The initial focus of the interagency committee has been placed on developing proposed standards for tabulation of data by business size and on developing approaches to resolving such problems as the difficulty of obtaining size data based on comparable reporting units and comparable indicators of business size.

Some promising approaches to improving small business data are being tried. IRS, for example, is currently linking payroll tax reports to corporate income tax returns in order to add employment and payroll measures to its corporate tax data base. And plans are underway to use various tax records to develop a longitudinal data base for a sample of business units. Nevertheless, the problems associated with improving the utilization of existing record collection mechanisms are formidable. One critical problem is the lack of adequate access to a systematic business list, such as the SSEL, which can be used to identify the various kinds of business reporting units and link together business reports in ways that desired variables can be tabulated on the basis of common size classifications and reporting unit concepts. Indeed, the SSEL would appear to be a central factor in efforts to solve a variety of data problems extending well beyond the need for small business data per se, and even involving a variety of problems relating to developing data files pertaining to individual workers. Because of its wide-ranging importance, the SSEL program is described in some detail in the next chapter. Issues of access to the SSEL are covered in Chapter VIII.

G. Appendix IV.1. Data from IRS and SSA

This appendix contains descriptions of (1) IRS administrative record data files; (2) special data files produced for the Bureau of the Census from IRS administrative files; (3) IRS sample data files developed from administrative records for statistical use; and (4) IRS Statistics of Income publications. In addition the appendix contains a list of data tables available in the Annual Statistical Supplement to the Social Security Bulletin.

1. Data from IRS

Administrative Record Data Files

Business Master File (BMF)--Contains selected data from the

return of partnerships, corporations, fiduciaries, charitable trusts, and business related data of exempt organizations. In addition, it includes data from a=, gift, and various excise tax returns, and employment tax return data is on this file for all entities.

Individual Master File (IMF)--Contains selected data from the tax return records of all individual income tax return filers including sole proprietorship data reported on Form 1040 Schedules C and F.

Exempt Organization Master File (EOMF)--Contains selected data from the return of exemptions which have been granted tax exemptions as organizations organized and operated exclusively for religious, charitable, educational, governmental, or similar purposes. This file is an information file whose primary function is to provide data to monitor the numerous types of exempt organizations. The organization is established on the EOMF when it applies for and is granted a tax exemption.

Employee Plans Master File (EPMF) - is maintained for use by the Internal Revenue Service, Department of Labor, and Pension Benefit Guaranty Corporation. The file contains selected data on

plan characteristics obtained from applications for plan approval or determination letters and data from the annual return records. Unlike the ODMF which only established an entity on the file when an exemption is granted, an entity is established on the EPMF upon receipt of an application for approval or determination letter, or when an annual return is filed.

Individual Retirement Arrangement File (IRAF)--Contains selected data on individual retirement arrangements. Special Data Files Produced for the Bureau of the Census from Master Files.

The Business Master File Entity Change File--this file changes and supplements the annual BMF. Changes are to entity name and address and filing requirements. New entities are added and indicators are set to mark inactive records.

Employer's Quarterly Federal Tax Return File--this file contains quarterly payroll, taxable tips and FICA wages paid for all companies with a 941 (domestic payroll), 941 PR (Puerto Rico payroll) and 941 SS (Virgin Islands, Guam, etc.) filing requirement.

Corporation and Partnership Return File--file contains large corporation (1120) and small corporation (1120S), and partnership (form 1065) annual receipts data.

Sole Proprietor Name and Address File - file contains names and addresses for sole proprietors who report profit or loss from business or profession (schedule C) and/or report farm income and expenses (schedule F).

1040 Schedule C and 1040 Schedule F Data File - this file contains receipts data and physical address for sole proprietor businesses.

Exempt Organization Business Income Information Return Files (990C, 990T, 990PF)--file contains business receipts for selected organizations exempt from filing an income tax return.

Employer's Annual Tax Return for Agricultural Employees File--file contains annual FICA payroll for all employers with a 943.

Alphabetic BMF Microfilm File (Name Directory)--this file is the Business Master File, in alphabetic sequence, on microfilm.

Sample Data Files for Statistical Use

Corporation Source Book--is based on a sample of corporation returns. It provides corporate income and balance sheet tables, by asset size for approximately 175 industry groups. These are available to the public for a charge on hard copy, on microfilm, and magnetic tape. These tables form the basis for the annually published reports, Statistics of Income, Corporation Income Tax Returns.

Statistics of Income Tape--derived from samples of United States individual, corporation, fiduciary, estate, partnership, exempt organization and pension plan returns are retained on magnetic tape. On a cost reimbursable basis, bona fide researchers may obtain copies of these tapes devoid of identifying and geographic information.

Individual, Proprietorship, Partnership, and Corporation Tax

Model File--files which are based on the Statistics of Income samples, and are available annually, contain, in general, the data present in our annual individual Corporation and Business Income Tax Returns reports. On a reimbursable basis, the Service will general statistical tabulations or simulate the administrative and revenue impact of law changes. The identity of taxpayers is kept confidential in these files. For individuals, proprietorships, and partnerships, the most de-

Annual Statistics
Periodic of Income
and Supplemental Publications
Reports

Department of the Treasury

Internal Revenue Service

Publication 711 (Rev. 7-80)

Publications are for sale by the Superintendent of Documents,

U.S. Government Printing Office, Washington, D.C., 20402

Other Reports Periodic and as Supplements

Estate Tax Returns, 1976

Publication 764

Gross estate by type of property

Lifetime transfers by asset type

Funeral and administrative expenses

Other deductions

Taxable estate; Estate tax, Tax credits

Data classified by- Taxable and nontaxable status; Size of gross

estate; Estate valuation method; Size of net

worth; Age, sex and marital status of decedent;

Tax rates; States

Personal Wealth Estimated from Estate Tax Returns, 1972

Publication 482

Provide estimates of the asset holdings if the living population

with gross wealth of more than $80,000:

Composition of assets

Distribution of assets by age, sex, and marital status.

Number of millionaires by three measures of wealth

Distributions by value of corporate stock, and by value of

real estate.

Historical statistics, selected years.

Fiduciary Income Tax Returns, 1974

Publication 808

Sources of Income, Taxable income

Exemption and Deductions

Income tax and tax credits

Additional tax for tax preferences; Allocation of accumulation distributions

Data classified by-

Trusts and Estates; Tax rates and type of tax; Size of total Income

Historical statistics, selected years

Sales of Capital Assets Reported on Individual Income Tax Returns, 1973

Publication 458 (scheduled September 1980)

Number of transactions

Gross Sales price

Cost or other basis plus expense of sale

Gross gain or loss

Details on sales of residences

Details on sales of business and farm property

Data classified by- Type of asset; Short-term vs. long-term; Length of period held; Taxpayers age 65 and over; States; Size of adjusted gross income; Size of net capital gain or loss.

Individual Retirement Arrangements, 1976

Publication 1107 (scheduled August 1980)

Number of arrangements

Contributions

Compensation

Distributions

Penalty taxes

Data Classified by- Type of arrangement; Source of Compensation; Size of adjusted gross income.

Private Foundations Exempt From Income Tax, 1974

Publication 1073 (scheduled September 1980)

Receipts, Including contributions, gifts, and grants

Deductions

Net Income

Net Investment Income and Tax

Assets and Liabilities]

Minimum Investment return

Distribution amount

Qualifying distributions]

Undistributed Income

Excise taxes paid by foundations

Unrelated business Income and tax

Data classified by- Exempt activity; Accounting period, State

Size of- Total receipts, Net income; Total assets

Small Area Data From Individual Income Tax Returns, 1974

Publication 1008

Number of returns and exemptions

Adjusted gross income

Salaries and wages

Dividends in adjusted gross income

Interest received

Total tax

Data classified by- Metropolitan areas; Counties; States; Size of adjusted gross income

(Report for 1978 scheduled December 1981)

International Income and Taxes, Domestic International Sales Corporation Returns, 1972-1974

Publication 1071

Receipts, including qualified export receipts

Deductions, including export promotion expenses

Net income

Amounts deemed or actually distributed

Assets and liabilities- Trade receivables; Producer loans; Capital accounts by type

Gross receipts of the DISC

Current and prior year gross receipts of the DISC and related U.S. persons

Data classified by- Country of destination; Product; Industry; Accounting period

Size of- Total gross receipts; Total assets of both DISC and corporate parent

(Report for 1975 scheduled December 1980)

International Income and Taxes, Foreign Tax Credit Claimed in Corporation Income Tax Returns, 1968-1972

Publication 479

Foreign tax credit-

- Foreign income and taxes
- U.S. net income and tax
- Data classified by-
  - 1968 and 1972; Foreign country; U.S. industry
- Credit limitation method
- Size of-

Total assets; Foreign tax credit; U.S. net income

1969 and 1970; Total assets; U.S. Industry

Western Hemisphere Trade Corporations, 1968 and 1972

(Report for 1974 scheduled September 1980)

Data similar to those for 1968 and 1972 for corporations with assets of $250 million or more

International Income and Taxes, U.S. Corporations and their Controlled Foreign Corporations, 1968 and 1972

Publication 1026

Net income and tax of U.S. parent corporations

Earnings, tax and transactions by type of foreign corporation with U.S. parent corporation and other related persons

Data classified by- Foreign country, Year of incorporation, Size of total assets, industry, and accounting period of both U.S. parent and foreign corporation

(Report for 1974 scheduled February 1981)

Data similar to those for 1968 and 1972 for corporations with total assets of $250 million or more

International Income and Taxes, Foreign Income and Taxes Reported on Individual Income Tax Returns, 1975

Publication 1100

(scheduled October 1980)

Exemption if income earned abroad-

Income earned abroad for personal services

Tax-exempt amount

U.S. taxable income and tax

Data classified by- Foreign Country; Type of residence status abroad; Size of adjusted gross income

Foreign tax credit-

Foreign income and taxes

U.S. taxable income and tax

Data classified by Foreign Country-

Credit ??? method; Size of adjusted gross income

Annual Statistics of Income Complete Reports

Individual Income Tax Returns Publication 79

Presents Information

annually or periodically on-

Sources of Income, including-

Salaries and wages

Dividends; Interest

Rents and royalties

Business or profession

- Farm
- Capital gains; Ordinary gains
- Pensions and annuities

Adjusted gross Income

Adjustments to Income

Exemptions

Computation of Itemized deductions, including-

- Contributions; Medical
- State and local taxes paid
- Home mortgage and total Interest paid

Zero bracket amount (standard deduction)

Taxable income

Income tax

Maximum tax

Tax credits, Including-

- Child care credit
- Earned Income credit
- Foreign tax credit
- Investment credit
- Jobs credit

Residential energy and business energy Investment credits

Retirement Income credit

Minimum tax and tax preference items

Tax withhold or due at filing time

Payments of estimated tax

Tax overpayment credits and refunds

High Income returns

Data classified by-

Size of adjusted gross income

States

Tax rates and type of tax computation

Taxpayer marital status

Taxable and nontaxable status

Tax payers age 65 or over

(Report for 1978 scheduled November 1980)

Corporation Income Tax Returns Publication 16

Presents Information annually or periodically on-

Receipts, including-

- Business receipts; Capital gains
- Rents and royalties
- Domestic and foreign dividends
- Taxable and nontaxable Interest

Deductions, including-

- Cost of sales and operations
- Advertising; Rents; Repairs
- Interest and taxes
- Employee benefit plans
- Depreciation, depletion, and amortization

Depreciation under ADR procedures

Net Income and taxable Income

Statutory special deductions

Income tax

Foreign tax credit

Investment credit

Work Incentive credit

U.S. possessions tax credit

Minimum tax and tax preference items

Tax payments and overpayments

Distributions to stockholders

Book vs. tax net income

Consolidated returns

Small Business Corporations

Domestic International Sales

    Corporation returns

Members of controlled corporate groups

Foreign corporations with U.S. business operations

Foreign owned U.S. corporations

Number of pension plans

Assets and liabilities-

    Notes and accounts receivable

    Investments in Government obligations

    Depreciable and depletable assets

    Accounts payable

Mortgages, notes, bonds payable

Net worth

Data classified by-

Industry; Accounting period

Returns with net Income

Size of- Total assets; Income taxed at normal and surtax rates; Business receipts; Income tax

(Report for 1978 scheduled February 1981)

Business Income Tax Returns Publication 438

Sole Proprietorships and Partnerships

Presents Information annually or periodically on-

Number of-

Sole proprietorships

Partnerships; Partners

Receipts, Including-

- Business receipts
- Partnerships-
  - Dividends; Interest
  - Rents, Royalties

Deductions Including-

- Cost of sales and operations
- Interest and taxes
- Rents; Repairs
- Depreciation, depletion, and amortization

Net Income

Profitable businesses

Inventories

Payroll

Partnership payments to partners

Partnership payments to retirement plans

Depreciation under ADR procedures

Cost of depreciable property

Partnership capital gains

Sale proprietors' adjusted gross Income and source of nonbusiness

Income

Partnership assets and Liabilities

Limited Partnerships

Jobs credit computation

Investment credit computation

Business energy investment

credit computation

Partnership tax preference Items

Date classified by-

Industry; State

Number of partners

Number of retirement plans

Partnership year of organization

Partnership accounting period

Sex of sole proprietor

Size of- Receipts; Partnership assets; Sole proprietorship

net income; Sole proprietors adjusted gross income

(Report for 1977 scheduled November 1980)

Preliminary Reports Precede complete report - contain several basic tables

Individual Income Tax Returns, 1979

Publication 198 (scheduled February 1981)

Corporation Income Tax Returns, 1977

Publication 159 (scheduled November 1980)

Business Income Tax Return, 1979

Publication 453 (scheduled November 1980)

Reports currently available Use order form provided on back

tailed data we could produce would be by Internal Revenue Service District. In most cases, districts are geographically coterminous with States; however, there are four districts in New York State, and two each in Pennsylvania, Ohio, Illinois, Texas, and California. We do not publish geographic data for corporations since the place where the return was filed may be different from the location of the principal business activity.

Statistics of Income Publications

Statistics of Income publications include annual reports based on individual corporate, and business returns; occasional reports based on other tax returns and schedules; and Supplemental reports classifying information from individual returns by geographic areas (SMSA and county) prepared biennially. Among the occasional reports are:

Fiduciary Income Tar Returns--this report presents estimates

of total income and its composition, deductions, taxable estate, and tax for personal trusts with income $600 or more for which a fiduciary flied an income tax return, Form 104 1. Important classifications include type of trust, size of total income, and tax rate.

Estate Tar Returns--this report presents estimates of gross estate by type of property, deductions, taxable estate, and tax for decedents with gross estate in excess of $60,000 for whom an executor filed an estate tax return, Form 706. Important classifications include size of estate, tax rate, and State.

Personal Wealth Estimated from Estate Tax Returns--this report presents estimates of the number and wealth of that portion of the population with assets of more that $60,000 based on the application of mortality weighting factors to estate tax return data. Important classifications include age, sex.marital status, as well as various measures of gross and net wealth.

Sales of Capital Assets reported on Individual Income Tax Returns--this report presents estimates of the transactions by type of property, gross sales price. basis of property and expense of sale, and net gain or loss reported on individual income tax re-

turns with sales of capital assets. Important classifications include size of income including and excluding capital gain or loss. and size of net gain or loss.

Returns of Private Foundations Exempt from Income Tax--this report presents estimates of the receipts, expenditures, net income, assets and liabilities of organizations classified as private foundations (and exempt from income tax) which file Forms 990-PF. Additional data are provided on excise taxes relating to excess investment income, investments jeopardizing exempt purpose, and prohibited expenditures.

Farmers' Cooperative Income Tax Returns--this report presents estimates of the receipts, deductions, net income, tax, assets, and liabilities for both exempt and nonexempt farmers' marketing and purchasing cooperatives filing on Form 990-C and 1120, respectively. Important classifications include type of service, type of commodity marketed, and State.

Returns of Employees' Pension Plans and Pension Trusts--this report presents estimates of the receipts, disbursements, assets and liabilities of individuals or organizations who maintain employees' pension plans or pension trusts and who file an annual

statement on Form 4848, 4849, and 990-P. Additional data include type of entity, type of plan, method of funding, and number of employees covered and not covered.

Returns of Organizations Exempt from Income Tax--this report presents-estimates of the receipts, expenditures, assets and liabilities of organizations (other than private foundations) exempt from income tax under Section 501 (c) of the Internal Revenue Code and which file Form 990. Important classifications include the subsection of the Code under which exempt and the principal business activity.

The description of available Statistics of Income reports on pages 36 and 37 is copied from recent SSI publications.

2.. Data From SSA

The following pages list data tables published by SSA in its Annual Statistical Supplement to the Social Security Bulletin. The list is copied from the Supplement which presents data for 1976. SSA's sample data files maintained in connection with the

Continuous Work History Sample program are described in Appendix III.2.

List of Tables'

Table No.		Page No.
	General	
	Social Security and the economy	
1.	Gross national product and social welfare expenditures under public programs, fiscal years 1928-29 to 1974-76	44
2.	Social welfare expenditures from public funds in relation to total government expenditures and Federal grants to State and	

local governments, fiscal years 1928-29 to 1974-76

3. Public programs: Social welfare expenditures, fiscal years 1928-29 to 1974-76 45

4. Aggregate and per capita national health expenditures, by source of foods and percent of gross national product, fiscal years 1929-76 46

5. Amount and percentage distribution of personal health care expenditures for the aged, by type of expenditure and source of foods, fiscal year 1976 46

6. Personal income and social security payments, 1929-76 47

7. Labor force and estimated workers covered under social insurance programs, 1939-76 48

8. Total earnings, wages and salaries and earnings in employment covered by selected social insurance programs, 1946-76 49

Poverty data

9. Weighted average poverty thresholds for non-farm families, by size, 1959-77 50

10. Trends in poverty: Number and percent of persons poor, by age, 1969-76 51

11. Trends in poverty among families: Families in poverty, by sex, age, and work experience of head, 1959-76 52

12. Poverty status and current living arrangements of persons aged 65 and over 52

13. Poverty status and work experience of family heads and unrelated individuals, by age and sex 53

14. Poverty states of aged households receiving social security benefits 54

15. History of Federal minimum wage rates under the Fair Labor Standards Act, 1938-79 55

Interprogram social security data

16. Social insurance and veterans' programs: Cash benefits and beneficiaries, by risk and program, 1940-76 56

17. Veterans' programs: Veterans receiving compensation or pension, by type of payment, and age, 1940-76 58

18. Selected social insurance and veterans' programs: Benefits, by State 59

19. OASDHI and selected public assistance programs: Average monthly payments in current and 1975 prices, 1950-76 60

20. Rejected social insurance programs: Source of funds from contributions and government transfers, 1965-76 61

21. Selected social insurance trust funds: Financial operations, 1937-76 62

22. Unemployment trust food: Status, 1940-76 63

23. OASDHI and SSI: Population aged 65 and over receiving OASDHI cash benefits, SSI payments, or both, 1940-76, and took by State,1976 64

24. Federal grants: Total to State said local governments, by purpose, fiscal yews 1929-30 to 1974-76 65

25. Federal grants: Total to State and local governments, amount and percent, by purpose and by State (ranked), fiscal year 1976 67

26. Unemployment insurance: Summary data on State programs,1940-76, and by State, 1976 68

27. Temporary disability insurance: Selected data on State and

railroad programs 69

28. Workmen's compensation: Coverage,benefits, and costs, 1940-76 70

Food stamp program

29. Number of persons participating, value of bonus coupons,and average bonus per person, 1962-76 71

Old-Age, Survivors, Disability, and Health Insurance

Trust funds

30. Old-age and survivors insurance trust fund: Status, 1937-76 72

31. Disability insurance trust fund: Status,1957-76 73

32. Combined OASI and DI trust funds: Status, 1957-76 74

33. Hospital insurance trust fund: Status, 1966-76 75

34. Supplementary medical insurance trust funds: Status, 1966-76 75

Workers

35. Workers,earnings, social security numbers issued, and employers reporting taxable wages under OASDHI, 1937-76 76

36. Workers and earnings of wage and sooty and self-employed workers, 1951-76 77

37, Farm workers under OASDHI, 1951-75 78

38. With taxable earnings, by type of worker and sex, 1937-76 79

39. With taxable earnings (all and 4-quarter): Percent with annual earnings below taxable limit, by sex, 1937-76 80

40. With taxable earnings: Number, by age and sex, 1937-76 81

41. With taxable earnings: Median earnings, by age and sex, 1937-76 82

42. With taxable wages (all and 4-quarter): Number, by wage interval, 1937-76 83

43. With taxable wages (male, all and 4-quarter): Number, by wage interval, 1937-76 84

44. With taxable wages (female,all and 4-quarter): Number, by wage interval, 1937-76 85

45. With taxable earnings (self-employed): Number,
by age and sex, 1951-76 86

46. With earnings credits (self-employed): Number,
by earnings-credits interval and sex, 1951-76 87

47. With taxable earnings: Number,earnings,and contributions,
by type of employment and State 89

48. Insured: By insured status, 1940-77 90

49. Insured: By insured status, sex,and age, 1972-77 91

50. Insured (aged 65 and over): Number eligible for and
percent receiving benefits, by sex and age, 1941-77 92

51. Insured (aged 62 and ever): Number eligible for and percent
receiving benefits, by sex and broad age group, 1956-77 93

Summary benefit date

52. Total benefits paid, by type of program,1937-76 94

53. Number and average,monthly benefits in current
payment status, by selected family groups,1940-76 95

54. Benefits in current-payment status, number and amount,

by type of beneficiary, 1940-76 96

Benefits awarded

55. Individuals: By type of beneficiary, 1940-76 97

Social Security Bulletin, Annual Statistical Supplement, 1976 1

Table Page

No. No.

56. Conversions: Number and average monthly amount, by reason

for conversion,type of benefit awarded, and previous

type of benefit 98

57. Retired workers: By states of award and sex, 1950-76 99

58. Retired workers with and without reduction for early retirement: Number ant! average amount, by status of award and sex, 1956-76 99

59. Retired workers with and without reduction for early retirement: Number and percent, by monthly amount and sex 101

60. Retired workers with and without reduction for early retirement: Number and percent, by primary insurance amount and sex 102

61. Retired workers,disabled workers,and widows: Average amount and, for retired workers, primary insurance amount, 1940-76 103

62. Disabled workers: By monthly amount and sex 104

63. Wives and husbands: By type of beneficiary, 1950-76 105

64. Children: By type of child beneficiary, 1940-76 106

65. Mothers: By type of mother beneficiary, 1950-76 107

66. Widows and widowers: By type of entitlement, 1950-76 107

67. Lump sum and survivor, Workers represented and average payment, by type of award, 1940-76 108

Benefits awarded and/or In current-payment status

68. Individuals: By type of beneficiary, race, age, and sex 109

69. Individuals: Number and average amount, by type of beneficiary, alt, sex, and race 120

70. Women beneficiaries: Number and average amount, by type of beneficiary and race 122

71. Individuals with reduction for early retirement: Number and average amount, by type of beneficiary, race, age, and sex 123

72. Wives with reduction for early retirement: Number and percent, by type, 1956-76 126

Benefits In current-payment states

73. Individuals: Number and average age, by type of beneficiary 127

74. Individuals: Number and average amount,

by type of beneficiary and race 127

75. Aged beneficiaries, By age, sex and race 127

76. Aged beneficiaries: By type of beneficiary,
age, and sex 128

77. Retired workers with delayed retirement, credit: Number,
average amount, and average primary insurance amount,
by age and sex 128

78. Retired workers without reduction for early retirement and
without delayed retirement credit: Number and
average monthly amount, by sex and age 129

79. Retired workers: Number and percent, by year of
entitlement and sex, 1940-76 130

80. Disabled workers: Number and percent, by year of
entitlement and sex, 1960-76 131

81. Widows: Number and percent by year of entitlement,
1940-76 131

82. Retired workers and dependents: Average amount,
by type of beneficiary and sex, 1940-76 132

83. Retired workers: Number, average age, and percent,
by sex and age, 1940-76 132

84. Retired workers and dependents: Number and percent, by type of beneficiary and primary insurance amount 134

85. Retired workers with and without reduction for early retirement: Number and percent, by monthly amount and sex 134

86. Retired workers with and without reduction for early retirement: Number and percent, by primary insurance amount and sex 135

87. Retired workers with benefits in nonpayment status; Number and percent, by monthly amount and sex 136

88. Dual entitlement: Persons with retired-worker and secondary benefit, with and without reduction for early retirement, by primary insurance amount and sex 137

89. Dual entitlement: Persons with retired-worker and secondary benefit, by type of secondary benefit and sex, 1952-76 137

90. Retired workers with and without reduction for early retirement: Number and average amount, by sex, 1956-76 138

91. Retired workers: Percent, by monthly amount, age, and sex 139

92. Disabled workers and dependents; Number and percent, by type of beneficiary and primary insurance amount 140

93. Disabled workers: Number and percent, by monthly amount

and sex 141

94. Disabled workers and dependents: Average benefit, by type of beneficiary, 1937-76 142

95. Disabled workers: Number and monthly amount, by sex, 1957-76 142

96. Disabled workers: Number, average age, and percent, by age and sex, 1957-76 143

97. Wives and husbands: Number and monthly amount, by type of beneficiary, 1950-76 144

98. Children: Number and monthly amount, by type of child beneficiary, 1940-76 145

99. Children: Number, by type of child beneficiary and sex of worker, 1950-76 146

100. Survivors of deceased workers: Average amount, by type of beneficiary, 1940-76 147

101. Survivors of deceased workers: Number and percent, by type of beneficiary and primary insurance amount 147

102. Mothers: Number and monthly amount, by type of mother beneficiary, 1950-76 148

103. Widows and widowers: Number and monthly amount, by basis of entitlement, 1950-76 149

104. Retired-worker, survivor, and disabled-worker families: Number, average primary insurance amount, and average benefit, by family group 149

105. Retired-worker, survivor and disabled-worker, families: Number, average primary insurance amount, and average amount payable, by family group with special minimum benefit 151

106. Disabled-child families: Number, average primary insurance amount, and average amount payable, by family group 151

107. Student-child families: Number, average primary insurance amount, and average amount payable, by family group 152

108. Retired-worker and disabled-worker families: Percent, by monthly amount 153

109. Survivor families: Percent, by monthly amount 154

Benefits withheld and terminated

110. Withheld from individuals: Number, by reason and by type and age of beneficiary 155

111. Withheld from wives and husbands and from children: Number, by

reason and type of beneficiary 155

112. Workers' compensation offset for disabled worker families: Number and average amount before and after onset, by type of family 156

113. Terminated for individuals: Number, by type of beneficiary, 1940-76 156

114. Terminated for individuals: Number, by reason and type of beneficiary 157

115. Terminated for wives and husbands and for children: Number, by reason and type of beneficiary 157

Table Page

No. No.

Benefits paid

116. Total paid from OASI trust fund: Amount and percent,
by type of beneficiary, 1940-76 158

117. Total paid from DI trust fund: Amount and percent,
by type of beneficiary, 1957-76 159

State monthly benefit date

118. Cash benefits paid: Total, by program 160

119. Benefits in current-payment status: Number,
by type of beneficiary 161

120. Benefits in current-payment status: Amount,
by type of beneficiary 162

121. Benefits in current-payment status: Number,
by age, race, and sex 163

122. Retired-worker benefits in current-payment status:

Number and percent receiving, by monthly amount, ranked by State average 164

123. Disabled-worker benefits in current-payment status: Number and percent receiving, by monthly amount, ranked by State average 165

124. Child benefits in current-payment status: Number, by type of child beneficiary and basis of entitlement 166

125. Retired-worker benefits in current-payment status: Number and average amount, 1940-76 167

126. Widow and widower benefits in current-payment status: Number and percent receiving, by monthly amount, ranked by State average 168

Beneficiaries residing abroad

127. Benefits in current-payment status: Number and total monthly amount, by country and type of beneficiary 169

Worker disability awards

128. Number and Percent, by selected causes of disability, 1957-74, and by sex, 1974 171

129. Diagnostic group: Number and percent, by age and race, 1974 172

130. Occupational division: Number and percent, by sex and race, 1974 173

131. Age on birthday in year of onset of disability: Number and percent, by sex, 1974 174

MEDICARE benefits

132. Hospital and supplementary medical insurance: Number of enrollees aged 65 and over, by age, sex, race, and census region, 1966-76 175

133. Hospital and supplementary medical insurance: Number of disabled enrollees under age 65, by age, sex, race, and census region, 1973-76 176

134. Hospital insurance: Number of enrollees,
by State, 1966-76 177
135. Hospital insurance: Number of bills approved for Payment and
amount reimbursed, by type of benefit and type
of beneficiary, 1966-76 178
136. Hospital insurance: Number of inpatient short-stay hospital
care bills, covered days of care, and charges, by type of
beneficiary, 1966-76 178
137. Hospital insurance: Average covered charge per covered day of
care in short-stay hospitals and skilled-nursing facilities,
by State, 1971-76 179
138. Supplementary medical insurance: Number of reimbursed bills,
charges and amount reimbursed,
by type of service, 1966-76 180
139. Supplementary medical insurance: Number of bills received by
carriers and assignment rates, 1969-76 181
140. Supplementary medical insurance: Reasonable charge
determination for claims assigned and unassigned, 1971-76 181
141. Hospital and supplementary medical insurance:
Benefit payment amounts, by State, 1972-76 182

142. Hospital insurance: Number of inpatient hospital and skilled-nursing facility admissions and rates per 1.000 enrollees, by type of beneficiary, 1966-76 183

143. Hospital insurance: Number of inpatient hospital and skilled-nursing facility admissions and rates per 1,000 enrollees, by State and type of beneficiary 184

144. Hospital and supplementary medical insurance: Number of facilities and beds for participating hospitals, skilled-nursing facilities, home health agencies, and independent laboratories, 1967-76 184

145. Hospital insurance: Number of participating hospitals and beds per 1.000 enrollees, by State 185

146. Hospital and supplementary medical insurance: Number of participating skilled-nursing facilities, home health agencies, independent laboratories, and end-stage renal disease facilities, by State 186

Supplemental Security Income

147. Number receiving federally administered payments, and total amount, by reason for eligibility and State 187

148. Number receiving State-administered supplementation and total amount, by reason for eligibility and State 188

149. Number receiving federally administered payments and average amount, by reason for eligibility and type of payment, December 1976 188

150. Number of all persons receiving federally administered payments and average amount, by State, December 1976 189

151. Number of aged receiving federally administered payments and average amount, by State, December 1976 190

152. Number of blind receiving federally administered payments and average amount, by State, December 1976 191

153. Number of disabled receiving federally administered payments and average amount, by State, December 1976 192

154. Number and percent receiving federally administered payments, by reason for eligibility and living arrangements, December 1976 193

155. Number of adult units and children receiving federally administered payments and average amount, by type of payment and reason for eligibility, December 1976 193

156. Total payments, Federal SSI payments, and State supplementation, by State 194

157. Number of blind and disabled children receiving federally administered payments, by State 194

158. Persons receiving federally administered payments and number and percent in concurrent receipt of income, by reason for eligibility, source of income, and average amount, December 1976 195

159. Percent of persons in concurrent receipt of federally administered SSI payments and social security benefits in December 1976 and average amount of social security benefits, by reason for eligibility and State 196

160. Number and percent of all, persons receiving federally administered payments, by reason for eligibility, sex, and race, December 1976 197

161. Number and percent of all adults receiving federally administered payments, by reason for eligibility and age, December 1976 197

162. Number and percent of blind and disabled children receiving federally administered payments, by age, December 1976 197

Table No. | Page No.

163. Number and percent of persons receiving federally administered payments with representative payees, by team for eligibility, December 1976 197

164. Number and persons of individuals receiving Federal SSI payments, by reason for eligibility and monthly amounts December 1976 197

165. Number and percent of couples receiving Federal SSI

payments, by reason for eligibility and monthly amount, December 1976 197

Black Lung Benefits

166. Currently payable to miners, widows, NW dependents: Number and amount, 1970-76 198

167. Currently payable to miners, widows, and dependents: Number and monthly amount, by State 199

Public Assistance

168. AFDC and emergency assistance: Average monthly number of recipients, total amount of cash payments, and average monthly payment, 1936-76 200

169. OAA, AS, and APT, Average monthly number of recipients, total amount of cash payments, and average monthly payment, 1936-76 201

170. General assistance: Average monthly number of recipients, low amount of cash payments, and overall monthly payment,

1936-76 202

171. Public assistance: Vendor payments for medical care, by program, 1951-76 203

172. AFDC and emergency assistance: Average monthly number of families and recipients of cash payments and total amount of payments, by State 204

4 Social Security Bulletin, Annual Statistical Supplement, 1976

CHAPTER V

Developments in Data from Business Establishment Reporting

Non-standardized concepts, definitions, and procedures used in

developing administrative record sets create serious difficulties for statistical uses. The potential for major new uses of administrative records may in fact be quite limited because of these problems and other Problems such as incomplete establishment reporting, poor timing, and confidentiality restrictions. there are, however, some new developments which present opportunities for improving the coordination and statistical use of key administrative record sets.

This chapter examines three evolving programs which illustrate the potential and problems associated with efforts to improve the statistical utilization of business reports obtained in connection with tax-related administrative data collection. The programs are the Census Bureau's development of the Standard Statistical Establishment List, the Social Security Administration's effort to adjust its data programs to new administrative procedures calling for annual (forms W-2 and W-3) rather than quarterly (form 941) employer reports of individual worker wages, and the Bureau of Labor Statistics' cooperative program with State Employment Security Agencies to make statistical use of records collected in connection with Unemployment Insurance payroll taxes.

The SSEL program represents an explicit attempt to identify the most useful definition of business establishment units for statistical analysis purposes, and to build "bridges," when necessary, between these statistical units and legal entities for which tax and other administrative reports are available. The SSEL not only is intended to facilitate more efficient direct use of administrative records for statistical purposes, but it also has been planned as a vehicle for coordinating statistical data collection efforts so that data collected from business in different programs can more easily be compared and integrated. In this connection the SSA and UI payroll tax programs represent particularly important administrative data collection programs, because both payroll tax programs have statistical components which involve requests for multiestablishment businesses to provide supplemental "establishment" information with their tax reports in order to permit tabulation of employment and payroll data by industry and geographic areas. A number of important advantages could be derived from better coordination of the SSA and UI establishment reporting plans with each other and the SSEL; but there are also a number of legal, institutional, and technical

obstacles to improved coordination. The discussion in this chapter and much of the remainder of the report (especially chapters VII and VIII) illustrates these potential advantages and the barriers to improvement in addition to describing applications of the data collected through current business establishment reporting procedures.

While the emphasis in this chapter is on information collected from businesses, both the SSA and UI payroll tax programs involve the collection of data (from businesses) pertaining to individual workers. In fact, the focus of SSA statistical use of payroll tax data has been the Continuous Work History Sample program which is organized explicitly around individual worker records. The UI program has been directed primarily toward utilizing aggregate establishment reports of employment and payroll, but a program to develop a Continuous Wage Benefit History sample is underway using individual worker records collected in connection with the UI program. Just as a general coordination of the SSA and UI establishment reporting plans with the SSEL program would provide important statistical advantages. so would coordination and linkage of the CWHS and CWBH individual record systems. This chapter does

not deal with such individual record linkage efforts, but Chapter VI provides several case studies illustrating the advantages and problems associated with efforts to link data from various individual record systems.

A. Standard Statistical Establishment List

There has been a long history of endorsement of the general principle that a centrally compiled list of firms and their establishments should be available for multiagency use in the conduct of statistical samples. Presently, each government statistical agency is responsible for compiling and maintaining the business register needed

for their particular statistical applications. The use of independently developed lists. with attendant differences in definition and coverage. seriously affects the comparability of the economic data provided by the various agencies. and also results in considerable duplication of effort and costs and increases in respondent reporting burden. Concerns such a-, these constitute a substantial part of the criticism of government statistical programs.

The Office of Federal Statistical Policy and Standards of the Department of Commerce (formerly Division of Statistical Policy of the Office of Management and Budget) has been a consistent advocate of a central list concept. Towards this end. in 1968. the Bureau of the Census was designated by OMB as the focal agency for the development. establishment and operation of such a directory (known as the Standard Statistical Establishment List--SSEL) on behalf of Federal statistical agencies. Funding for the project started in fiscal year 1972 with an operational Directory available covering data year 1974.

Construction of the SSEL was known to be technically feasible since the methodology had been followed previously in assembling

the economic censuses mailing list and in utilizing administrative data. Since the linkage among the principal source agencies. i.e.. Census, IRS., and SSA is the common usage of the Employer Identification Number by all three agencies. and using the establishment as the basic "building block" of the SSEL, it is possible to link together and identify the affiliation of parent companies. subsidiary firms, and their establishments throughout all phases of economic activity.

1. File Construction

The SSEL now consists of a central multi-purpose computerized name and address file of all known multiestablishment and single establishment employer firms in the United States. The systems design for computer processing is predicated on variable word-length record which permits additional information to be added as desired.

2. Multiestablishment Firms

Information for multiestablishment firms was initially derived from Census Bureau records. From the 1972 Economic Censuses, the necessary basic information had been assembled for the organizational units of all firms included in the economic censuses. All establishments of these firms were linked to the enterprise level and were identified by their individual SIC codes, physical locations. employment size codes, etc.; and all known domestic establishments of these multiunit firms were identified regardless of activity. This practice represented a departure from that of previous censuses where records were maintained only for establishments engaged in activities defined as within the scope of the economic censuses. Multiestablishment companies not covered by the economic censuses were identified in a two-stage survey. In November 1972, as part of the Economic Census processing. all legal entities with 50 or more employees were canvassed to determine their enterprise structure. Each legal entity was requested to list all companies it owned or controlled and the name and EI number of its controlling company, if any. Information was also requested on employment, kind of industrial activity, and number of business locations operated under that EI number. Detailed

listings of establishments were not requested in this survey since the major emphasis was to consolidate those legal entities into their correct enterprise structure. This operation was coordinated with the regular Economic Census processing to produce an integrated file. A similar survey was conducted in January 1974 covering calendar year 1973 to canvass smaller entities with 20-49 employees. In addition, 175,000 small out-of-scope companies (less than 20 employees) were canvassed in 1974 if classified in an activity changed by the 1972 SIC revision.

3. Single Establishment Firms

Approximately 80% of the universe of business establishments with one or more employees are single establishment firms represented by one EI number. For these establishments. the enterprise, legal entity and establishment are identical. For this reason, information for single establishment firms was derived from the administrative records of other government agencies since it would be difficult to justify the government and respondent cost involved in duplicating this information by direct survey contact.

The Business Master File of IRS served as the basic universe

file from which the single unit company listing was derived. This source provided company name. address, EI number and legal form of organization for all firms with one or more paid employees.

March 12 employment and the Standard Industrial Classification Code were obtained from the records of the Social Security Administration. The four quarters of payroll were obtained from IRS records.

In constructing the multiestablishment company file. the Census Bureau recorded the EI number of the entity owning the establishment in conjunction with the SSEL File Number. Matching these EI numbers of multiunit firms against the Business Master File (El file) and unduplicating, the residual list resulted in the establishment of the single unit file. Using these inputs, the SSEL became operational covering data year 1974 and is now used as

the mailing list source and sampling frame for all Census Bureau economic programs.

4. File Maintenance

The use of administrative records has played an integral part in creating, maintaining, and updating the SSEL file. During noncensus years, the single establishment file (approximately 4 million records) is updated solely from administrative records. New births are received monthly from, IRS and SSA with information on name and address, EI number, SIC code and legal form of organization code. Employment and payroll data are received quarterly. Geographic codes are assigned by Census from the address information received from IRS and SSA.

For multiestablishment firms, a company organization survey was undertaken to insure that the organizational structure of each company is updated at least once each year. This survey includes companies in scope of the Economic Censuses as well as out-of-scope companies covered in a special survey. Preprinted forms are sent

to each company. listing all establishments known to be operated by it including name and physical location of each establishment. The company is requested to update these listings and report March 12 employment, first quarter payroll and annual payroll by establishment location. The reported payroll is then compared to the IRS administrative payroll at the EI and company level, and discrepancies resolved. In addition, administrative record employment and payroll is used to impute nonmail or delinquent companies. Several working papers describing the SSEL system have been written (U.S. Bureau of the Census. 1979). Copies can be obtained from the Census Bureau. Because of the cost of annual maintenance, a complete file of zero employee cases is available only from each quinquennial Economic Census.

5. Confidentiality

Current legislative restrictions, including title 13 of the Census Act, do not permit the release of the SSEL to other agencies for statistical use. Legislation has been proposed, however, which would permit the release of this file to certain other Federal

agencies (see Chapter VIII).

B. W-2 and W-3 Records

Starting in 1979 with data for tax year 1978, a significant change took place in the method of reporting to the Social Security Administration the wages paid to employees by their employers. A single annual wage reporting system began under which forms W-2 are used as the report of individual employee wages for both social security and income tax purposes. This eliminates the quarterly reporting of a detailed listing of wages paid to each employee covered under social security. Employers still have to file quarterly reports containing wage and tax liability information with the Internal Revenue Service. State and local government employment is excluded from the annual reporting system.

Under the annual reporting system, forms W-2 along with transmittal forms W-3 (see Figure V.1) are received at one of four SSA Data Operations Centers where the material is examined for completeness and correspondence initiated with employers having incomplete shipments. After microfilming, the documents are

prepared for optical scanning or key-to-tape operations. The data on the output tapes are then transmitted to SSA's Central Office via telecommunications equipment. Here the data are merged with data from employers who submit their reports directly on magnetic tape and all the data are subjected to a series of computer balancing and validation operations. All validated earnings items, those taxable under the Federal Insurance Contributions Act as well as other earnings, are forwarded to IRS for processing for income tax purposes. Copies of the validated FICA items are retained by SSA to update the Summary Earnings Record for individual employees.

The new W-2, W-3. reporting system has a number of positive and negative implications for SSA's Continuous Work History Sample statistical programs. (See Chapter III for a description of the current CWHS system.) The most important positive features of the new annual reporting system are that for the first time SSA will have information on total wages paid to an individual, thus eliminating the need to estimate wages above the maximum that is taxable for social security purposes; and that initially the system will include information on employees not covered by social security as well as covered employees. Privacy and Tax Reform Act

questions, however, remain to. be resolved relating to the extent to which data for uncovered employees can be used for statistical purposes in the CWHS program.

On the negative side, there will no longer be data on individual earnings amounts by quarter. Also, there are preliminary indications that the items for statistical processing will not be available until sometime later than under the quarterly reporting system. There are also indications that the SSA's Establishment Reporting Plan could be adversely affected because of the nature of the reporting requirements for forms W-2 and W-3.

Another aspect of the new annual reporting system that has great statistical potential is the employee's address on the form W-2. These addresses could be coded to obtain residence geographic information. Unfortunately, present procedure does not call for SSA to capture this information

in any machine readable form. However, the possibility of retaining this information in the future is presently being

pursued.

The units which employers use for establishing summary (W-3) reports presently differ widely among employers under SSA's voluntary establishment reporting plan (see chapter VII). If employers were to use the establishment definitions and codes developed by the Census Bureau for its Standard Statistical Establishment List, the resulting file of W-3 forms would be immensely more useful for statistical purposes that is if the W-3 forms were collected with Census Bureau establishment codes and confidentiality problems restricting SSA-Census interchange of records were resolved, the SSEL could be used to code establishments by industry and geographic location (State, county, and possibly subcounty units), The resultant file could be used to provide tabulations of annual wage and salary income and employment by industry for detailed geographic units. Such tabulations could be used to improve a number of statistical programs, including BEA's State and local area personal income accounts and the Census Bureau's County Business Patterns program. In addition, the improved geographic coding for the individual records (W-2's) associated with the W-3's would improve the CWHS program and if

used in conjunction with W-2 (or other;

residence information, would permit the development of valuable intercensal commuting estimates for local areas. Currently, however, not only is vital SSA access to the SSEL limited by legislation, but there would appear to be substantial employer resistance to proposals that they report to SSA on the basis of SSEL establishment concepts (which frequently involve more detailed establishment reports than called for in SSA's voluntary establishment reporting plan).

C. Unemployment Insurance System

A case study of the statistical usefulness of administrative

records for establishments can be gleaned from the unemployment insurance system. This system was established as part of the Social Security Act of 1935 to serve as a countercyclical income maintenance program for offsetting losses in wage and salary income of the experienced work force. Initially, UI covered only employers in the private nonfarm economy with eight or more employees. Over the years, the system has been continuously expanded. In March 1978, over 90 percent of employed workers were covered by the State and Federal UI system.

In the UI system, a variety of administrative data is maintained. Three important data sets which serve as the primary source of statistical uses are discussed in this Chapter (see Figure V.2).

First of all, there is a master list of more than 4 million subject employers which contains the names and addresses of covered firms and both actuarial and statistical information. Secondly, information from the quarterly tax reports filed by employers is maintained. Finally, in all but 12 States, firms report the total wages paid to each employee during the quarter to determine an individual's eligibility and benefit amount when filing a UI claim.

1. Master List of Employers

State agencies collect and process certain statistical information to help provide standardization for reports and tabulations. Employers are assigned county and industry codes. Industrial activity is reviewed on a three-year cycle, and attempts are made at identifying multiestablishment employers and setting in place a mechanism for Supplemental reports of employment and wages by county and industry. The UI list is used by State agencies to draw samples in the Federal-State programs sponsored by BLS and operated by the States. A number of States also use the list to publish industrial directories. The lists are provided to the Bureau of Labor Statistics to use for sampling purposes under a pledge of confidentiality. BLS uses the lists to develop its UI Name and Address File which serves as a sampling frame for its directly collected surveys.

The UI Name and Address File has a number of drawbacks. Since it is derived from an administrative source, many of the refinements needed for sampling purposes are not present. For example, the major identifying field in the file is a UI account

number which is assigned independently by the various States. There is no unique way to identify firms or companies within a corporate structure across States. Also, identification of multiestablishment employers varies from State to State.

2. Employers' Quarterly Tax Report

Taxes are collected quarterly from subject employers by mailing each employer a tax form on which he reports the total wages paid to employees during the quarter, the amount of these wages that is subject to taxes, the taxes due, and the number of employees on the payrolls for the period that includes the twelfth of each month. The tax forms are due at the State agency 30 days after the end of the reference quarter. Multiestablishment employers are also mailed a statistical supplement with their tax report requesting a breakdown of the monthly employment and wage figures by reporting unit. Five months after the end of the quarter, State summaries in machine readable form are sent to BLS, Washington. Two summaries are required of each State: (1) Statewide by four-digit industry, and (2) counties by two-digit

industry. States that can provide four-digit industry by county, need only send one summary. These summaries are called ES-202 reports.

Many programs of the BLS and BEA rely on the ES-202 report's employment and wage data. Within BLS, the Current Employment Statistics, Labor Turnover Statistics, the Occupational Employment Statistics. Industry Projections, and Occupation Safety and Health Statistics programs are benchmarked to industrial employment data emanating from the ES-202 report. The BEA national income and personal income estimates rely heavily on the UI administrative data. In addition, personal income is used in formulas to allocate billions in Federal funds to State and local governments. At the local level the average wages of workers covered by UI are used to adjust the average annual wage payments allowed Comprehensive Employment and Training Act Public Service Employees. The State agencies also make substantial use of employment and wage data to assess the economic vitality of local labor markets in their labor market information programs. Practically every employment related statistic that is generated in the BLS-BEA-State employment agency enclave has the UI administrative as its base. The ES-202 report

has its limitations and problems. There is no set mechanism of quality control to assure that

all subject employers are reporting. There is no program of quality assurance for ascertaining the accuracy of data reported by employers on their tax reports. Statistical reports which are a by-product of an administrative program often receive a low priority. The statistical functions in producing the ES-202 report compete for basic UI program resources with tax collections. benefit payment. and research activities. Hence, many States cannot fully implement industry coding and multiestablishment "breakout" activities.

3. Individual Wage Records

In most States, the collection of the quarterly tax reports also involves an itemization of individual workers' wage payments identified by social security number. This data base provides a

rich source of information on an individual's earnings history. The Current Wage and Benefit History program of the U.S. Department of Labor is attempting to tap this data base to link earnings experience with workers' eligibility and receipt of UI benefits. Since each individual's earnings are linked to the employer, studies on wage dispersions by industry and county (on a place of work basis) are feasible. These files are also being used to map mobility patterns and labor turnover actions as part of Labor's Employment Service Potential program.

4. Improving Data Quality

The UI administrative data have room for improvement because of the large and cumbersome task of identifying multiestablishment employers. Their major strength is the quarterly collection and timeliness versus other sources of establishment records-namely, the Census Bureau's County Business Patterns program. Census does considerable work annually in identifying and maintaining multiestablishment breakdowns of firms in its Company Organization Survey. Access to these data could help identify and refine multiestablishment reporting problems in the UI record system.

At the same time, one of the weaknesses of the Census' establishment records is the industry codes of single-establishment firms. Those single unit firms not covered in the 1972 or 1977 Economic Censuses retain industry codes assigned from information submitted when the application for an EI number was made. A matching of industry codes in the two data system could improve the coding of single establishment firms on the Standard Statistical Establishment List and help identify potential problem areas between the two systems; i.e., such a match could determine how much of the difference between BLS and Census series is due to coding, how much is due to reporting differences, and how much is the result of differences in treatment of central administrative Offices.

CHAPTER VI

Potential Uses of Administrative Records for Data Linkages: Selected Case Studies

A. Introduction

In this chapter case studies of ongoing or completed research using administrative records for data linkage studies are compiled. These studies are in various stages of development; some have been completed, others are in the planning stages, and still others have been partially implemented. Nevertheless, each included study serves to illustrate important aspects of the research potential and problems associated with uses of administrative records.

The individual case studies exemplify the potential uses of administrative records for linkages, illustrating some of the benefits derived and the difficulties involved. The wide range of

general issues addressed include confidentiality concerns, operational feasibility, and data quality. The specific topics discussed are the data sources and identifiers used for matching, the criteria used to determine acceptable matches, and methods used to improve the quality of identifiers. Project goals, and the general methodologies used to carry out the match will also be discussed for these selected cases.

Administrative records have been used in the past in a number of interagency data linkages for statistical purposes. For example, matching studies involving record checks have been conducted to evaluate the last three decennial censuses. Although the case studies presented in this chapter differ in scope, methods and objectives, they serve to illustrate some of the ways administrative records can be used for statistical purposes:

1. The Linked Administrative Statistical Sample (LASS) project is an effort to produce an improved data base for mortality research by integrating samples from the record systems of three agencies: IRS, NCHS, and SSA.
2. The Use of Administrative Records in the Survey of Income and Program Participation (SIPP) illustrates the use of

administrative records in multiple frame Surveys, where issues of sampling efficiency are central, and in response error studies where the validity of survey reported data are compared to program data. Future use of administrative records in the SIPP will emphasize data base enhancement through the integration of difficult to collect data obtained from administrative records with survey collected data.

3. Use of IRS/SSA/HCFA Administrative Files for 1980 Census Coverage Evaluation describes a multiple systems estimation procedure which will be used to obtain estimates of Census coverage for States and selected subgroups of the population.

4. Record Linkage in the Nonhousehold Sources Program is a study to improve the coverage of the 1980 Census in which administrative data sets (drivers license records and Immigration and Naturalization legal alien records) are used to augment the information in another data set (1980 Census enumeration records).

References to published and unpublished material related to the study are included at the end of each case study. The

supplementary information may provide interested readers with more detail on the studies themselves and on the difficulties in successfully implementing the linking of data files.

In all administrative matching studies, conceptual differences and operational difficulties, including access to administrative records, may impede or even invalidate the attempt. However, the analytic potential of obtaining an expanded, more detailed data base through successful matching is so great that complicated and careful procedures are often worth the effort. The increasing numbers of attempts to improve statistics through matching testifies to this conclusion.

B. Case Study 1: Linked Administrative Statistical Sample (LASS) Project

The Linked Administrative Statistical Sample or LASS project is an, effort to upgrade the Social Security Administration's Continuous Work History Sample. The primary focus of the study is to examine the issues surrounding the development of integrated

samples from the record

system of three agencies: the Internal Revenue Service, the National Center for Health Statistics, and the Social Security Administration. The principle objective of the project is to create an improved data base for mortality research.

The material presented here discusses a few of the major concerns which are being addressed in order to determine the feasibility of producing such a sample. Organizationally this case study is divided into two main parts. The first of these sets the background of the study, its research objectives and the specific data sources to be included. The second describes the initial planning activities being engaged in and some of the progress which has been made thus far in each area. There are also some

concluding comments on the issues to be faced if the project is to enter an operational phase.

1. Background and Initial Project Goals

For over 40 years [1] both government and nongovernment researchers have made extensive use of statistical information about American workers derived from the Continuous Work History Sample (CWHS). The primary Social Security use made of the CWHS has been in tabulating the characteristics of covered workers to keep track of how this group has changed over time with changes in the Social Security Act and in the demographic mix of the population [e.g., 2]. The Bureau of Economic Analysis has made considerable use of the CWHS as a source of regional workforce characteristics and especially changes in the workforce, both geographical and industrial [3]. Uses by nongovernment researchers have also been extensive, covering the gamut from labor market supply questions to the measurement of lifecycle earnings [e.g., 4-5]. Recently in a pioneering effort by Goldsmith and Hirschberg [6] attention has been focused on the CWHS' potential to address

industrial and environmental health issues.

While the usefulness of the CWHS data has been demonstrated repeatedly, it is limited in scope, content, and quality by program requirements. These weaknesses would, of course, have to be corrected in order for the files to reach their full potential as a general purpose data base for statistical research. The support of present and potential users who recognize the importance of these data will be necessary to bring about the changes which will improve its usefulness [7].

Professionals concerned with epidemiological problems, occupational safety, and general environmental issues are among those interested in an improved, augmented CWHS. In fact, the real start of the Linked Administrative Statistical Sample project was a meeting at the National Center for Health Statistics (NCHS) in October of 1978 involving representatives of several agencies. including Social Security, to explore areas of mutual concern that relate to epidemiology studies.

When the U.S. Congress [8] amended the Public Health Service Act (Public Law 95-623), NCHS's mission for conducting and coordinating research activities aimed at improving all aspects of health services in the United States was greatly broadened. Part

of this legislation calls for the development of a plan by the National Center for Health Statistics for the collection and coordination of statistical and epidemiological data on the effects of the environment on health. Therefore, NCHS desired to work with other agencies to find feasible, cost-effective approaches to developing an implementation plan for carrying out its new mandate.

One effective and relatively inexpensive way to achieve this goal is to integrate data already collected by Federal agencies in pursuit of their individual missions. Social Security and the Internal Revenue Service (IRS) are two of the major agencies which have current data that are not generally available for epidemiological studies. The proposed LASS project is an attempt to exploit these data systems for studying the occupational and industrial etiology of disease.

a. LASS data elements

The Linked Administrative Statistical Sample is to retain the same simplicity of design as the CWHS, and takes that sample as its starting point. In particular, it is planned that ultimately a

common statistical sample will be created which is based on the ending digits of the social security numbers used to select the one percent Continuous Work History Sample. The following data elements are proposed for inclusion in the final linked sample:

1. Mortality information from the National Center for Health Statistics' processing of death certificates. (At a minimum, on a prospective basis the fact of death would be confirmed by matching the National Death Index to the CWHS. Also, the basic demographic items from NCHS's statistical record including cause of death would be added. Retrospectively, similar information might be obtained as far back as the late 1960's for every identified CWHS decedent. Finally, for both the retrospective and prospective efforts, the decedent's usual occupation and industry during his or her lifetime, items not now coded by NCHS, would be obtained from the certificates themselves.)

2. Individual income tax items obtained initially from the Statistics of Income (SOI) program. Eventually, the information will be derived directly as a by-product of

processing. (Detailed income, deduction and tax data could be obtained from the Transaction Files now used to update the Master File. Also available from that source would be any neeD residence information. Last, but not least, the occupation entry on the return would have to be transcribed to the statistical records.)

3. Longitudinal earnings and benefit histories developed at Social Security as part of the Continuous Work History Sample. (The CWHS, as it now exists, can provide basic demographic information for the sampled individuals, details on every covered job by industry and place of work since 1956; total covered earnings since 1936 (by

year since 1950); and, for beneficiaries, the nature of their claims and the amounts they and their dependents receive in benefits.)

b. LASS research goals

There are a number of general Ion run goals of the LASS effort. Three major ones are listed below:

1. To develop a basic source of socioeconomic and job-related mortality and morbidity data. The resulting statistical sample proposed hem could be used to construct mortality rates by age. race, sex, industry, occupation, and place of work or residence. This could lead eventually to a much greater understanding of the etiological factors associated with cancer and other causes of death. By following individuals over time by occupation, industry and residence, for example, it may be possible to separate out the effects of these factors on health from the effect of health on these factors.

2. To construct longitudinal personal and administrative unit income profiles of the population at the National,

State, and Substate regional levels. These income distributions could be studied both before and after the imposition of Federal income and payroll taxes.

3. To study regional labor market conditions using the data on industry, occupation, wages, and self-employment earnings along with basic demographic characteristics such as age, race, and sex. Mobility studies and other such work now done with the CWHS [3] would be greatly enhanced by the augmented dataset available under this proposal. Particularly important in this regard is the occupation and residence data that might be obtained from tax returns. (For workers who don't file tax returns, residence information will be available from the new annual wage reporting system based on the W-2.)

The short-run goals of the project are centered around feasibility questions such as assessing data quality and estimating operating costs. An examination of a few of these goals is provided in the next section along with a summary of the work done so far to achieve them.

2. Pilot Activities and Feasibility issues

In planning for the operational phase of the LASS project a number of activities have been undertaken. Included among these are-

1. attempting to resolve the confidentiality concerns of the participating agencies,
2. examining coverage and content differences between SSA and NCHS death information,
3. determining the problems which arise when adding cause of death and other data from death certificates to the CWHS,
4. assessing the codability and validity of the occupation entry on the individual income tax return,
5. developing procedures for upgrading the CWHS data on industry and place of work, and
6. studying the completeness of the W-2 residence information.

Full details on the progress to date may be found in the LASS Working Notes Series [9] or in the publication Statistical Uses of

Administrative Records with Emphasis on Mortality and Disability Research [10]. In what follows, only a brief overview has been given.

a. Resolving privacy concerns

Many privacy concerns must be addressed before the LASS project becomes operational. in addition to disclosure laws with government-wide application such as the Privacy and Freedom of Information Acts, each of the participating agencies has legal constraints--statutes and regulations-which control access to its microdata. At minimum, these need to be coordinated in terms of some unifying principles of interagency data sharing. In addition, some of there may need to be amended. For example. the Tax Reform Act of 1976 has changed the character of information from earnings reports for persons who are in covered employment under the Social Security Act by defining this as tax return information subject to confidentiality restrictions in the Internal Revenue Code [11]. The Act allows IRS to disclose identifiable tax return data to SSA only if those data are required for the operation of SSA programs

or for IRS tax enforcement purposes. These conditions will almost certainly be too restrictive for some of the activities planned for the CWHS. if the interpretation IRS has given the Tax Reform Act prevails

[12], corrective legislation may be needed to overcome these problems.

Privacy requirements also raise policy issues. Should projects involving the linkage of records from various agencies be undertaken at all if there is any identifiable future possibility that the resulting data will be used in form for administrative or enforcement Purposes?

SSA protects linked statistical files from non-statistical use by regulation, but this may not have the force and permanence afforded by the "shield" laws protecting Census Bureau and NCHS

data. and possibly also IRS data. On the other hand, these statutory confidentiality shields also circumscribe the development and use of linked data in identifiable form outside each respective agency, even for statistical purposes. In the short-term pilot phases of this work, the confidential data contributed by NCHS could be protected by making SSA staff "Special agents" or temporary employees of NCHS.

Such a procedure has worked well in past linkage studies (e.g., the 1973 CPS-IRS-SSA Exact Match Study [13]); nonetheless, a firmer basis is needed before this project reaches its operational phase. i.e., by FY 1982 if not sooner.

Discussions among the participating agencies to address the many privacy issues are still at a fairly early stage. Legislative initiatives we proceeding, in order to protect SSA data and to resolve problems of making tax return information available for statistical linkage. Various Presidential proposals aimed at providing government-wide legislation for protection of statistical and research data offer a major step towards resolving the access issues raised by this project.

Given the potential for disclosure that this rich data base

would have, the creation of public use files from an upgraded CWHS presents difficulties which, at present seem insurmountable. To service potential users, we have been considering the possibility of setting up a Research Center that would provide tabulations and other statistical summaries. Computerized methods such as random rounding routines [14], would be built into such a center's procedures so that the possibility of any inadvertent disclosures could be prevented. (it is anticipated that such a center could be largely user supported.)

b. Examining SSA-NCHS death reporting differences

There are two key questions that must be answered if the SSA death reporting system is to be used to study industrial mortality differentials:

1. How complete is the reporting of deaths to SSA?
2. Are there differences in the information shown on death certificates and SSA records?

The reporting of deaths to Social Security is not required for

persons who are not OASDI beneficiaries; however, financial incentives, like the lump sum death benefit. make such reports common practice. In order to determine the characteristics of persons whose deaths are not "captured" by SSA, a cooperative project--the 1975 NCI-NCHS-SSA Mortality Study--was initiated with the National Center for Health Statistics and the National Cancer Institute (NCI); this study took as its starting point a sample of 23,000 deaths reported to NCHS for 1975. To date SSA has obtained the death certificates of these decedents and has nearly finished matching there to agency records. A paper presenting preliminary results were given at the August 1979 meetings of the American Statistical Association [15]. Present plans call for the coverage (or completeness) check to be followed by a comparison of the agreement between conceptually identical items like age, race, sex, and place of birth.

c. Adding data from death certificates to the CWHS

To add cause of death to the CWHS it is necessary to supply each State with lists of the decedents identified using SSA

information on name, social security number, race, sex, date of birth and date of death. Each State vital records office will then have to search its (microfilm) files and send copies of the death certificates to Social Security.

Several unanswered questions exist about this fairly simple process. Among these are

1. Will all the States be able to cooperate?
2. Will SSA's information be sufficient for the States to attempt a search?
3. What will be the quality of the searching?
4. What will be the total cost in money, time and staff?

A pilot test is now underway which should help address these questions. Information on every decedent in the CWHS who was identified as dying in 1975 has been sent to the States for death certificate searching. The CWHS decedents were combined, before being sent, with a subsample of NCHS cases already returned as part of the 1975 NCI-NCHS-SSA Mortality Study. Merging the two sets of decedents so they are simultaneously searched will make it possible to measure the quality of the work done in each State. (the NCHS cases were previously located by the States using death certificate numbers; now they will be located using SSA identifying information

which does not include the certificate number).

d. Usability of IRS occupation information

For a number of years there has been a continuing (and growing) interest among professionals concerned with epidemiological problems, occupational safety and general environmental issues, etc., in augmenting the

Continuous Work History Sample with an occupational variable. One approach for obtaining occupational data for earners in the CWHS is to use the information from returns. This creates difficult individual income tax problems given the uncertainty of the inclusion Of the OccuPation item on the tax return from Year to

Year as well as the lack of taxpayer instructions for reporting occupation.

One of the activities undertaken in preparation for the LASS effort was to compile the many studies [16] which have been done of the reporting of occupation on tax returns in order to make the call that this very important Content item be transcribed routinely as part of the Statistics of Income (SOI) program. The evidence from these studies suggests that at the major group level IRS occupation data may be roughly comparable in quality to that in the decennial censuses [17].

As part of their Statistics of Income Tax Year 1979 program, IRS has agreed to pick up occupation information. This effort will be supported by SSA with the ultimate objective of determining the feasibility and cost Of coding occupations for the entire set of tax returns in the 1 percent CWHS.

At present a collaborative pilot study of the SOI procedures is now underway involving a systematic sample of 6,700 returns. Some results from this pilot will be available in 1980. Plans for validating the occupation entries obtained in the Statistics of Income program are also being developed.

e. Upgrading CWHS industry and Place of work data

One of the most important parts of the LASS effort is to upgrade the quality of the CWHS coding of industry and place of work. To this end, there must be a further strengthening of the existing cooperative efforts between the Bureau of Economic Analysis (BEA) and. SSA in thoroughly examining the data quality problems which exist in the CWHS [7]. Equally important is the need to revitalize and expand the longstanding cooperative arrangements between the Census Bureau and SSA.

With respect to the BEA-SSA relationship, at present, plans call for the development of a detailed set of procedures to "perfect" the CWHS files for the period 1957/1977. A comprehensive approach to the handling of misreported (and/or missing) data is anticipated from this joint BEA-SSA effort. The data editing and imputation tasks are expected to be quite formidable indeed. Because of their one-time nature, the use of an outside contractor seems advisable (assuming the Tax Reform Act is changed to allow it). If all goes well the RFP could be written by FY 1982 with the

work potentially taking place during 1982-84. Joint BEA-SSA plans are also being developed to handle the new (post 1977) data quality problems that are being encountered in the changeover to annual wage reporting.

It is also expected that the Census Bureau will participate in the CWHS upgrading. This effort, however, will have a different focus from the plans developing with the Bureau of Economic Analysis. Traditionally, the Social Security Administration has provided industry and Place of work data for new employers to the Census Bureau in connection with the Bureau's Standard Statistical Establishment List (SSEL) program [18]. After each Economic Census the Bureau has returned to the Social Security Administration updated industry data for use in the CWHS. For single establishment employers the incorporation of this data in the CWHS is fairly routine. For multi-unit employers real difficulties arise because of differences in the identification of establishments between Census and SSA plus, of course, failures by SSA to obtain establishment-level information from some employers at all.

Two major changes in this arrangement are being proposed: (1) that the Bureau provide to SSA from the SSEL annual updates on

place of work codes for single-unit employers (again if the confidentiality issues can be worked out); and (2) that for multi-unit employers an experimental study be undertaken to see if the SSEL information on employer place of work can be combined with the employee's residential address (from the individual income tax return or the W-2) in order to create synthetic establishment identification codes for CWHS cases where the voluntary SSA establishment reporting plan is not working properly.

The synthetic establishment assignment process, as it is envisioned to date would use a Census Bureau address coding scheme to determine the distance between the employee's home and all the establishments of his employer. The establishment closest to the employee's residence could be chosen as the establishment that was "most likely" to be the employee's place of work. Complications caused by address changes over time would have to be overcome; but the scheme, in our opinion, offers real promise and should be tested. It is important to point out that discussions with the Census Bureau on these recommendations are at a very early stage. Realistically the likelihood is low that much progress will be made on this effort during 1980 or even 1981. However, some parts of

the task can be carried out during the period, e.g., coding the addresses of the employees. Building the full-scale system envisioned here would probably have to take place starting in 1982 or later.

f. Evaluating W-2 residence data

One of the advantages of the switch to annual reporting is that it provides access to new information not available

in the old quarterly system. The residence data from the Form W-2 is perhaps the most important new item, however, for cost reasons (and because of the complications inherent in the conversion). the W-2 residence data is not being processed for administrative

purposes. A pilot effort is now underway, though, to determine the usability of this data for statistical purposes. In the pilot, an attempt is being made to go back to microfilm copies of the original source documents from the employers. Microprints will be made and then examined for legibility and completeness. If the address data proves adequate, the W-2 could be a valuable adjunct to the IRS tax returns as a source of residence information for the CWHS. Consideration also will be given to using the W-2 addresses in a mail survey to learn about the occupation of income tax nonfilers.

3. Operational Implementation Issues

In order to mount the proposed Linked Administrative Statistical Sample project, a high degree of cooperation is essential both within Social Security's Office of Research and Statistics and among the other agencies involved. Most of the technical problems which must be faced have already been touched on in this note. Perhaps the hardest problems to be faced, as in any large endeavor, are organizational or managerial in nature.

Although meetings with both the potential producer and user agencies have been held frequently since October 1978, the LASS project is still in its initial planning phase. It will be some time before all the options have been laid out and the costs estimated. Establishing priorities will be a difficult process since each participating organization has its own missions. research goals and administrative procedures. There is also a concern about the ability of each of the participating agencies to obtain the new staff and budget that will be required.

Because of the formidable technical and resource problems that must be overcome, it is envisioned that a 5 to 10 year developmental period will be needed before the, Continuous Work History Sample can be used to its fullest potential as a vehicle for monitoring industrial and occupational health questions. In the interim, the intermediate products will be shared widely with interested members of the research community. To this end there was a special session at the 1979 Annual Meetings of the American Statistical Association on the LASS project [10]. Another such session is scheduled for the 1980 meetings.

For more information on the LASS program, contact:

Faye Aziz

Office of Research and Statistics

Social Security Administration

1875 Connecticut Avenue. N.W., Room 320H

Washington. D.C. 20009

Beth A. Kilss

Statistical Division PR:S

Internal Revenue Service

1201 E Street, N.W.. Room 403

Washington, D.C. 20224

4. References

[1] Buckler, W. and Smith. C., "The Continuous Work History Sample: Description and Contents," Policy Analysis with Social Security Research Files, U.S. Social Security Administration. 1978.

[2] U.S. Social Security Administration. Annual Statistical

Supplement series to the Social Security Bulletin.

[3] U.S. Bureau of Economic Analysis. Regional Work Force Characteristics and Migration Data (A Handbook on the Social Security Continuous Work History Sample and Its Application), 1976.

[4] Schiller, B., "Relative Earnings Mobility in the United States," Policy Analysis with Social Security Research Files, U.S. Social Security Administration, 1978.

[5] Jacobson, L., "Worker Displacement in the Steel Industry," Policy Analysis with Social Security Research Files, U.S. Social Security Administration. 1978.

[6] Goldsmith, J. and Hirschberg. D.. "Mortality and Industrial Employment (1)" J. Occupational Medicine Vol. 18. pp. 161-164, 1976. (Them were also two other papers by Goldsmith in this journal and an important letter commenting on the results by Pierre De Couffle.)

[7] Cartwright. D., "Major Geographic Limitations for CWHS Files and Prospects for Improvement." Review of Public Data Use. March 1979.

[8] Public Law 95-623, 95th Congress. 92 STAT. pp. 3443-3458.

[9] U.S. Social Security Administration. LASS Working Notes Series, Nos. 1-7. 1979.

[10] U.S. Social Security Administration. Statistical Uses of Administrative Records with Emphasis on Mortal and Disability Research (Selected papers

given at the 1979 Annual Meeting of the American Statistical Association in Washington, D.C.), October 1979.

[11] Alexander, L. and Jabine, T., "Access to Social Security Microdata Files for Research and Statistical Purposes." Social Security Bulletin, August 1978 .

[12] Alexander, L., with Scheuren, F. and Yohalem, M., "The 1976 Tax Reform Act arid the Statistical Program of the

Office of Research and Statistics," working paper prepared for the Subcommittee on Oversight of the House Ways and Means Committee.

[13] Kilss, B. and Scheuren, F., "The 1973 CPS-IRS-SSA Exact Match Study," Social Security Bulletin, October 1978.

[14] Fellegi, I. P. and Phillips, J. L., "Statistical Confidentiality: Some Theory and Applications to Data Dissemination," Annals of Economic and Social Measurement, National Bureau of Economic Research, April 1974. For a more recent and complete discussion see Statistical Working Paper No. 2: Report on Statistical Disclosure and Disclosure Avoidance Techniques, Office of Federal Statistical Policy and Standards, 1978.

[15] Alvey, W. and Aziz, F.. "Mortality Reporting in SSA Linked Data: Preliminary Results," Social Security Bulletin, November 1979.

[16] U.S. Social Security Administration, LASS Working Notes No. 2, January 30, 1979.

[17] Koteen, G. and Grayson, P., "Quality of Occupation Information on Tax Returns," 1979 American Statistical Association Proceedings.

[18] U.S. Bureau of the Census, Standard Statistical Establishment List program, Technical paper No. 44, January 1979.

C. Case Study 2: The Use of Administrative Records in the Survey of Income and Program Participation

The Office of the Assistant Secretary for Planning and Evaluation within the Department of Health and Human Services (HHS), in cooperation with the Bureau of the Census, initiated a joint statistical project called the Survey Of Income and Program Participation (SIPP). A fundamental objective of the SIPP is to provide data to support policy analysis of a wide range of Federal transfer and service programs. The survey data will be used to analyze the Impact of Federal programs, to estimate program participation and eligibility fates, future Program costs and coverage, and to assess the effects of alternative policy decisions on the various programs. Timely estimates of participation will be

provided for many existing programs, as well as estimates of the joint receipt of benefits across several programs. The survey will also support separate analyses of characteristics of Persons and families who an eligible but not participating in specific programs. When possible, survey data will be supplemented by administrative record data.

in addition to collecting program and eligibility data, the survey is expected to produce, on a timely basis, a comprehensive assessment of the economic circumstances of the population. The assessment is intended to cover objective factors (e.g. income, wealth, employment and family status) and selected subjective measures (e.g., attitudes and expectations about programs and personal well-being). The assessment will provide repeated observations on the same individual to permit the measurement and analysis of change over time. To supplement the analytical program undertaken primarily by the Department of Health and Human Services and the Bureau of the Census, a series of public use data The will be distributed at cost to researchers outside the government. These tapes should provide a rich and, in many ways, unique data base for studies of the working of government programs, the

economy, and society at large. The field activities have been undertaken to examine and resolve content, operational, and technical issues prior to beginning the ongoing SIPP in 1982:

1. Site Research.--a small experimental study of 2,800 households in five locations designed to provide a formal test of alternative survey design features, specifically recall period and questionnaire format.
2. 1978 Panel.--a national survey of 2,400 households designed to evaluate the implementation of a number of field and processing activities.
3. 1979 Panel.--a national survey of 11,000 households designed to study the effects of, (1) alternative questionnaires on income recipiency, (2) self vs. proxy response, and (3) length of recall on property income data.

A characteristic of the sample design common to each field activity was the use of several sample frame for the

selection of survey respondents. The frames which were used included a general area frame and special list frames consisting of administrative records from several HHS programs. Probability samples were drawn independently from each frame. Subsequent to each field activity. sample survey records were matched to their corresponding administrative records. Although there has been continuity of the learning process concerning matching and the use of administrative records within the developmental stages of the SIPP. the objectives of each matching operation have varied somewhat as the program has developed.

1. Objectives and Description

a. Site research

In the Site Research Survey, administrative records were used as sampling frames primarily to facilitate evaluation of

the experiments on alternative survey design factors. Two program recipient files were used as list frames in addition to a general area frame in each of the five locations. The first file was the June 1977 Aid to Families With Dependent Children (AFDC) master file maintained by the Texas State Department of Public Welfare in Austin, Texas. This file is an administrative system which maintains data on benefit amounts, payment history, demographic characteristics, and other information needed to administer the program. The second program recipient file used for selecting persons in the Site Research Survey was the Supplemental Security Record (SSR) maintained by the Social Security Administration (SSA) in Baltimore. This record is the national master administrative file for data on Supplemental Security income (SSI) benefits amounts, payment history, and demographic data. Table 1 provides an indication of how the sample households were distributed among the sample frames and questionnaire types in the Site Research. Table 2 exhibits the number of completed adult interviews for each sample frame.

A match of survey records to administrative cases drawn from each file was initiated to determine the accuracy and quality of

the income data collected in the Site Research. By comparing survey data to the record (control) data, the match allowed a validation and response error analysis with subsequent evaluation of the effects of the experimental treatments on income reporting. Because data on income types other than AFDC or SSI was not current and of questionable quality. the analysis comparing the survey data with administrative data was possible only for the AFDC or SSI income.

The Statistical Methods Division (SMD) of the Bureau of the Census was responsible for defining sampling specifications for the three samples and for drawing the area probability sample. The two samples from program data were selected by the respective agencies according to specifications developed by SMD.

Because of the small sample sizes and limited geographic scope, no effort was made to develop multiple frame estimates for the Site Research Survey. Thus, the file matching task was relatively uncomplicated; only the cases drawn from each record system needed to be matched with their respective survey records. The general population sample was not part of the matching operation.

The variables used to identify a match depended on the

availability of information in the administrative record system. Since each sampled address was assigned a unique control number. the matching of administrative records to their respective survey records involved essentially a two-stage process. First. the control numbers of the sample addresses were matched to the survey records. Then, within each household on the matched household file, a person match was attempted using the Social Security Number (SSN) of the individual on the administrative record as the primary match variable. Difficulties with matching on SSN at the person level were resolved by using age and sex as discriminating variables.

Although the process was used for both administrative record systems, an essential difference existed between the SSI match and the AFDC match. In the case of the SSI match, a manual match. using the procedure defined above, preceded an automated match. Since the., cases which could not be matched manually were discarded, the automated match appeared to be perfect. An exact count of discarded SSI cases is not readily available, but the Census Bureau has indicated with some assurance that there were relatively few.

An automated procedure which mirrored the manual procedure described above was used for matching the AFDC sample survey data with administrative record data. Descriptive statistics for the entire Site Research file are not available; however, a sampling of the results of the match procedure is given in Table 3. It is the authors' understanding that these results are representative of the results of the entire AFDC matching operation.

One rather disturbing finding of the Site Research matching procedure was that in a large number of cases (up to 30 percent), the individual selected from the administrative record system was not included in the household roster at the address shown in the survey record. This resulted from the procedures used to identify the sample unit. Interviewers were instructed to locate the sample address (which was not always found) and interview the residents in the household. They were not told to search for the specific individual on-the administrative record system, because of the fear that such a procedure would bias the survey data.

b. 1978 panel

In the second phase of the SIPP developmental field work, the

1978 Panel, a nationwide area probability sample of 1,950 households and a list sample of 411 households drawn from SSI files was interviewed at quarterly interviews over a period of 15 months. The purpose of again including this frame was to continue the investigation of SSI reporting with new survey techniques. Some of these techniques affected the general quality of all income data (e.g. new interviewer training procedures); other techniques were specifically developed to improve SSI reporting (e.g. distinguishing the color of the government-issued checks). Although no experiments were involved, the matching of survey data to administrative records has proved most informative in the evaluation of these new techniques. Thus, the context of the matching

Table VI.2.3. A Sampling of AFDC Matching Results

in the Site Research Survey

November ISDP-3

130	records to be matched
15	records matched Non-interview Households
73	records matched on HHLD ID. and SSN
9	records matched on HHLD ID.. sex and age
33	records matched on HHLD ID.. but could not match at person's level

October ISDP-4

127	records to be matched
12	records matched Non-Interview Households
74	records matched on HHLD ID. and SSN
12	records matched on HHLD ID.. sex and age
29	records matched on HHLD ID.. but could nix match at Person's level

January ISDP-10

III records to be matched

3 records matched Non-interview Households

71 records matched on HHLD ID. and SSN

8 records matched on HHLD ID.. sex and age

29 records matched on HHLD ID.. but could not match at person's level

January ISDP-15

129 records to be matched

15 records matched Non-interview Households

69 records matched on HHLD ID. and SSN

13 records matched on HHLD ID.. sex mid age

32 records matched on HHLD ID.. but could not match at person's level

Note: Data concerning the ISDP-20 and ISDP-25 Questionnaires

are not readily available at this time: however, according to the Demographic Surveys Division of the Bureau of the Census. the results of these matching operations are similar to the ISDP-10 and ISDP-15 match results.

activity, once again, has been limited to response error and validation studies for one specific income type. Preliminary efforts at multiple frame estimation in the 1978 Panel were considered in the early planning stages. However, because of the small size and low precision of the sample, this was not pursued.

The goals of the 1978 Panel matching operation did not substantially differ from the goals of the Site Research. Some refinements in locating list frame sample respondents and in the matching procedures resulted in a higher match rate in the 1978 Panel than had occurred in the Site Research Survey. To insure that the list frame person was a member of the interviewed households interviewers were instructed to go to the address listed and ask for the person (by name) selected from the administrative record. The interviewers did not know how these people had been selected; they only knew that the survey respondents were members

of a "person" sample rather than a "address" sample.

If the person did not live at the address. procedures were developed to assist the interviewing staff in locating the sample persons and interviewing there at their current address. These procedures were not always successful and some sample loss occurred when list frame persons could not be located.

Table 4 provides the results of the automated match of SSI data to the survey respondent for the 1978 Panel. The matching procedures for the 1978 Panel respondents were similar to those of the Site Research Survey. Unique household control numbers, assigned to each sample address at the time of sample selection, were used to match at the household level. Within the household, the sample program person was matched to his/her administrative record using the SSN as the primary match variable. Non-matches resulting from this procedure were clarified by comparing the age and sex variables. As can be seen from Table 4. data on the number of person level matches using only the Social Security Number are not available.

Table VI.2.4.--SSI Match Results for the 1978 Panel

April 1978 ISDP-303

486 records to be matched

1 record did not match at HHLD level

58 records matched at HHLD level only

427 records matched at person's level

July 1978 ISDP-403

491 records to be matched

23 records did not match at HHLD level

51 records matched at HHLD level only

417 records matched at person's level

October 1978 ISDP-503

496 records to be matched

29 records did not match at HHLD level

49 records matched at HHLD level only

418 records matched a person's level

January 1979 ISDP--603

496 records to be matched

30 records did not match at HHLD level

76 records matched at HHLD level only

390 records matched at person's level

April 1979 ISDP-703

497 records to be matched

38 records did not match at HHLD level

72 records matched at HHLD level only

387 records matched at Person's level

Note: An increase irk the number of household level non-matches occurred in interviews two through five because households which were not interviewed were included in the total number of

records to be matched. Since a questionnaire was not completed for these households, a non-match was assured.

c. 1979 panel

The use of two administrative record systems was incorporated into the survey design for the third phase of the SIPP developmental field work, the 1979 Panel. Supplementary samples of 1,000 program participants each were drawn from the December 1978 Supplemental Security Record file maintained by SSA in Baltimore and the 1978-1979 Basic Educational Opportunity Grants (BEOGs) applicant file. In the former, the respondents selected were blind and disabled SSI recipients; in the latter case, the applicant file was restricted to those determined eligible for a grant in the 1978-79 academic year. The 1979 Panel is still being fielded, and thus, no results are yet available regarding the use of the

different sampling frames. Plans, however, include the use of these administrative record systems both to evaluate income reporting and to obtain multiple frame estimates, thus improving the reliability of data regarding households in their programs. The latter goal will considerably complicate the matching process, since for this purpose the identification of the overlap domain among the three sampling frames (i.e., area sample, SSI, and BEOGs) is critical. The former goal is comparable to the work performed previously in the Site Research Survey and the 1978 Panel. That is, by matching administrative records to the survey records, detailed program data can be compared with interview data. Thus, further analyses and evaluation of the quality of SSI and BEOGs reporting are planned. Since this work is quite similar to the work completed on the 1978 Panel, the match of the sample individuals selected from administrative records with their survey records will follow essentially the same scheme as was used in the 1978 Panel. The improved field procedures for locating the list frame persons have been repeated, as well as the computer match process. In addition, the area sample will be matched to each of the administrative records universes, so that a study of reporting

of the respective income types for the area sample can be conducted.

The creation of multiple frame weights requires that sample respondents be placed in the correct domain of membership. Since the 1979 Panel is a developmental effort, two approaches to this problem will be compared:

1. asking the sample respondents questions to determine their domain membership; and
2. matching the individual survey records with the universes of the administrative records systems used for sampling.

In the first case, responses to survey questions about participation in the SSI and BEOGs programs are used as indicators of membership in the overlap domain. This approach may be particularly unsatisfactory because self-identification of membership in the exact universes in an interview tends to be very difficult. The BEOGs cases were drawn from certified eligibles, not all of whom are necessarily recipients; and, the SSI cases were drawn from blind and disabled, but not aged recipients. It is difficult to formulate appropriate questions to permit proper identification, and even more difficult for the respondent to give

an accurate response. Questions have been developed and will be asked in the Fifth Wave of the panel to permit determination of respondents' ability to self-identify membership in the programs. In the second case, a match of survey records for all interviewed individuals to both administrative universes is proposed. However, deficiencies in the quality of the matching variables, particularly the SSN and date of birth, will result in an undetermined number of false non-matches (i.e., a person interviewed in the sample survey who had non-zero probability of selection from one of the administrative lists, but whom the record match did not identify as having such a probability). In order to reduce (but not eliminate) false non-matches resulting from inaccurate or incomplete survey data, a set of procedures to validate and correct survey-reported SSN's or to supply missing SSN's will be implemented in conjunction with the Office of Research and Statistics (ORS)/SSA. Of course, these procedures will not assist the SIPP in determining false non-matches resulting from inaccurate data on the administrative record system.

The 1979 Panel matching procedures for the multiple frame domain determination have, at this time, yet to be defined. It is,

however, obvious that the SSN will be the primary matching variable with name, date of birth, race, age, and sex serving as confirmatory variables. The results of this exercise should provide valuable insight into the procedures required for a timely and operationally successful multiple frame sample survey.

2. Major Difficulties

The major problems arising from the use of administrative records for sampling have been consistent throughout the SIPP developmental work, affecting, to different degrees, all the sampling frames which have been used and/or considered in the program. The problems stem from difficulties in:

1. Identifying individual sampling unit with a known probability of selection;
2. Locating units in the sample in the field;
3. Gaining access to the administrative files;
4. Determining matches and non-matches;
5 Gaining timely access to updated administrative data for addition to the sample survey records; and

6. Finding administrative sources that are national in scope or similar from State to State.

The basis of the first problem lies in the fact that a one-to-one correspondence does not generally exist between survey units and the units on the administrative record files. A survey unit, in the SIPP developmental work, is a household. However, in many administrative record systems, the SSI system for example, a household can be identified by more than one individual's record (e.g. when more than one person in a household is a ,program participant). the is also possible that a single administrative record can lead to more than one household, such as when the record relates to a nuclear family which lives in more than one household. In the case of the SSI system, records were maintained in such a way that duplicate records for spouses could be deleted; records

for other recipients in the same household-were not unduplicated. In other programs, however, an unduplication process was not available or readily derivable, and sampling was deferred until such time as methods could be devised.

The second area of difficulty--identifying units selected in the sample in the field--was briefly mentioned in the section on the 1978 Panel. The problem primarily resulted from; 1) inadequate or inaccurate home addresses of individuals on the administrative records or 2) recent moves by program participants. In the latter case, a time lag of 2 to 4 months from sample selection to interview date contributed to the problem. These difficulties were handled by procedural changes in the 1978 and 1979 Panels, instructing interviewers to use the address only to locate the individual, and to interview the entire household where that individual was currently living. Individuals who moved and whose new address could not be determined remained a problem and could not be interviewed. In order to avoid violating the privacy of the sampled individuals and to avoid biasing the data, interviewers were only told that the "person samples". had been drawn from various government programs rather than from particular programs.

The problem of obtaining access to the administrative files which were ultimately used was not as difficult as had been anticipated. Most of the difficulty in this amp can be characterized as a substantial expenditure of staff time from the initial contact through the sample selection, and the production of a substantial amount of paperwork to obtain access to the administrative file. However, since the SIPP developmental work is a joint statistical project with the Census Bureau, confidentiality of the data being assured under Title 13, U. S. Code, access to the flies was granted.

Several brief, tentative efforts at using other program files maintained at the State or local level have been attempted. In these cases problems of access appeared more severe. The timing, amount of paperwork, and likelihood of being granted access dictated against vigorous pursuit of such frames during the early SIPP developmental program; further work in that area will be pursued later in the program.

The fourth problem of accurately identifying matches and non-matches between the survey records and administrative records was already discussed. The problem has not been resolved; however, the

experience gained from the Site Research and 1978 Panel has suggested that the quality of the survey data, particularly reporting of SSN, can be improved by emphasizing its importance in interviewer training. This, of course, cannot improve the quality of the SSN's on the administrative files. In the 1979 Panel, an attempt will be made to validate the SSN's provided by the respondents. Cases with invalid numbers will then be identified to the interviewers, in order that they may attempt to obtain a correct number during a later wave of the 1979 panel.

The type of matching operation conducted in the Site Research and 1978 Panel is considerably less sophisticated than that envisioned for the 1979 Panel. More will be learned during the next year concerning the SIPP's ability to match lists of survey respondents to administrative lists of program participants. The issues of survey reported and validated SSN's, inaccurate and incomplete data on the administrative file, and the use of multiple frame sampling in an ongoing survey will be affected by the ability to identify correct matches.

The fifth area mentioned--gaining timely access to updated administrative data to Supplement the sample survey records-my present a problem in the ongoing SIPP. In the SIPP program,

emphasis has been placed on providing relatively fast turnaround of the SIPP data for purposes of program evaluation and current assessments of the socio-economic well-being of the nation. If the sample design is dependent on access to administrative records for proper weighting, this access will have to be carefully timed to coincide with the end of data collection, or alternative means of providing preliminary data should be developed.

The last problem of finding administrative record sources that are national in scope or similar from State to State also will affect the ongoing SIPP. For many programs which have variable record systems at the State or local level, sampling may be operationally too difficult, despite the importance of the program. This will reduce the effectiveness of the survey in providing data on program participation for such programs.

3. Uses of the Administrative Files

Matched survey and administrative records of the Site Research Survey and the 1978 Panel data have not yet

been made available to the public since confidentiality issues still need to be resolved. In the case of the Site Research files, individual identifiers, such as SSN, address, name, and Census control number, have been removed and the income amounts above a fixed cutoff have been "topcoded" or reduced on the file to the cutoff point. Geographic codes identifying the city and the administrative record data remain on the file. These files are currently being edited and will be made available as public use tapes. Confidentiality issues will determine the final record layouts from this data collection activity.

Individual identifiers-that is, name and SSN--on the 1978 Panel quarterly tapes have been removed. However, at this time, administrative data, detailed geographic codes, and income amounts which have not been topcoded (i.e., coded to a fixed open-ended

category, usually $50,000 or more, if the amount exceeds the base of the open ended category) remain on the file. At this writing, all five waves of the 1978 Panel (unedited) have been received by HHS, from the Bureau of the Census. Current plans include making these tapes available as public use tapes once the confidentiality issues become resolved. Only two waves of the 1979 Panel have been received at this time. However, confidentiality issues concerning the administrative data should be resolved in a manner similar to the 1978 Panel data. The SIPP Staff intends to make these data available as public use tapes, retaining some minimal amount of information from the administrative records.

To date, the most important use of the matched files has been in the evaluation of reports of income recipiency. A major goal of the developmental work of the SIPP has been the improvement of reporting of income and related data through sampling procedures, questionnaires, and estimation techniques. The matching of survey reports to administrative records has allowed some objective evaluation of the efficacy of these efforts.

In the near future, the primary purpose of the use of administrative records for sampling will be to improve the

reliability of estimates of recipiency of relatively rare income types and of estimates of the characteristics of such recipients. Their income types will include both cash and non-cash transfers from federal programs. Through oversampling from program records and multiple frame estimation, the number of sample observations of program participants will be greatly increased, leading to improved reliability of program participation rates and characteristics of program participants. In addition, efforts will be made to add Social Security earnings records to the individual survey records, thereby enhancing the richness of the economic data base.

In the long run, administrative records may provide a means of adjusting SIPP estimates of recipiency and level of income using administrative control totals. Not as much thought has been given to this use as a means of developing better estimates of program participation. However, alternative means of improving survey data with administrative data are available and will be explored in the SIPP. For example, if the administrative data are known to be accurate, and if practical, reliable matching procedures can be developed, then individual data items on interview records might be adjusted. Alternatively, administrative data could be used as control totals for adjusting aggregate estimates of recipiency of

particular income types or participation in particular programs.

4. Quality of Results

At this time, the only analysis of the quality of the procedures and resulting information has been in the evaluation of the effectiveness of using different field procedures to locate sample cases. Within the next year an evaluation of the quality of the matching process will be conducted using the data from the 1979 Panel.

For more information on the use of administrative records in the development of the Survey of Income and Program Participation, contact:

Daniel Kasprzyk

Income Survey Development Program

SSA/ORS/ISDP

Room 322B, Universal North Bidding

1875 Connecticut Avenue NW.

Washington, DC 20009

5. Bibliography

L. Hausman, Characteristics of selected income-tested programs (May, 1977; 289 pp.)

H. Huang and D. Kasprzyk, An examination of the relative benefits of selected sample designs for the SIPP. ISDP Working Paper #5 (November, 1978: 29 pp.)

R. Kaluzny. 'Site Test analysis: characteristics of the data base (May, 1979; 53 pp.)

R. Kaluzny and John Scott Butler, The effect of instrument design on the reporting of AFDC and SSI income: a multinominal approach (March, 1980; 35 pp.)

B. Klein, Validating AFDC recipiency from the site research survey using a known sample of recipients (Forthcoming in the documentation of the Site Research Test)

C. Lininger, Ed. Survey of Income and Program Participation (SIPP) confidential report (August. 1979. 77 pp.)

B. Mahoney. M. Ycas, D. Kasprzyk, and H. Huang, Trade-offs in the collection of income, wealth, and program statistics

(June, 1978; 29 pp.)

J. Steinberg, Multiple frame sampling approach-general framework of alternative approaches (December, 1976: 18 pp.)

J. Steinberg, Multiple frame sampling approach--pro- posed design of a pilot test (February. 1977; 18 pp.)

S. Stephenson, Ed. Survey research issues workshop: proceedings (August. 1978; 274 pp.)

D. Vaughan, Errors in reporting Supplemental Security Income recipiency in a pilot household survey (August, 1978. 6 pp.)

M. Ycas, An introduction to the Income Survey Development Program (August, 1979; 31 pp.)

D. Case Study 3: Use of IRS/SSA/HCFA Administrative Files for 1980 Census Coverage Evaluation

1. Introduction

One of the major objectives of the 1980 Census Coverage Evaluation Program is to develop estimates of the coverage of population and housing in the census at the state and substate level. The Current Population Survey (CPS). which is conducted on a monthly basis, will provide these estimates. Persons listed in the CPS are matched on a one-to-one basis with the census listing of names in order to estimate census coverage error.

Special enumeration surveys were conducted as part of the 1950 and 1960 census evaluation programs. However, the results of these studies were considered not to be successful for providing accurate estimates of the undercount for certain subgroups of the population. One can conclude from these results that certain types of persons enumerated in the census are much easier to enumerate in the CPS than persons missed in the census. This bias is often referred to as "correlation bias;- a major objective of the 1980

CPS-census match will be to reduce this bias.

There are two means by which the Census Bureau hopes to reduce "correlation bias"'

1. By maintaining, as much as possible, independence of the CPS and the census.

2. By utilizing "independent" administrative files for purposes of improving the estimates of coverage error.

It is the latter process that will be primarily addressed in this case study. To the extent that a satisfactory match between the administrative files and the census and CPS can be achieved without impairing independence of the sample data, we should be able to obtain more accurate estimates of coverage error than were obtained in 1950 and 1960.

Two administrative flies are being considered: the IRS tax return file for persons aged 17 to 64 years of age and the Medicare file for persons 65 or over. Two research projects are being conducted to determine the feasibility of using the IRS and Medicare files and will be described in this case study. They involve matching the February 1978 CPS records to corresponding IRS and Medicare records.

It should be noted that to a great extent this program is still being developed. Thus, the projects described in this report could be subject to revision.

2. Objectives of the Program to Estimate the Census Undercount

The primary objective of the 1980 Census Coverage Evaluation Program is to develop estimates of the coverage of population and housing in the census. The estimates can be made using two different methods: Demographic analysis and survey estimates.

A. Demographic Analysis--The demographic method (demographic analysis) of census evaluation that will be used involves developing expected values for the population at the census date by the adjustment and combination of demographic data from sources essentially independent of the census being evaluated and comparing these expected values with the census counts. The particular method that is used for demographic subgroups depends on the nature of the available data. For ages under 45 in 1980, estimates will be developed on the basis of birth, death and immigration statistics. For ages over 65 aggregated medicare data will

provided the basis for estimates of coverage. For the remaining age groups an analysis of all censuses since 1880. along with death and immigration statistics, provides the basis for developing coverage estimates in 1980 (1).

Demographic analysis will provide national estimates of net census errors for age, sex and race groups. These estimates are measures of net error for age, sex, and race groups, combining coverage errors and errors of content. The demographic method is considered by Census staff to be more effective than a post-census sample survey for developing satisfactory estimates of net census errors at the national level for the total U.S. population. However, problems do exist with demographic analysis. the major one is the estimation of the number of undocumented aliens. At the present time. nd definitive methodology is available for including this segment of the population in the demographic estimates.

Demographic analysis will also provide are estimates of net census errors for broad age categories, by sex, and for white and black racial groups. However, it is questionable whether they will be better estimates due those produced from the CPS and to what extent they will be utilized.

B. Current Population Survey (CPS)-The data does not currently exist for using demographic analysis techniques to provide reliable estimates of coverage error for subnational geographic area such as cities, SMSA's and revenue sharing areas; in addition, the data now available for demographic analysis cannot provide estimates of coverage error for some important socioeconomic categories. The Census Bureau will utilize the April, 1980 and the August, 1980 Current Population Surveys to fill this void. Persons listed in the CPS are matched on a one-to-one basis with the census listing of names. Census resources exist for providing reliable estimates of net coverage error at the state level for the total population. Furthermore, the CPS will enable methodology to be developed (e.g., regression-synthetic estimation techniques) that might provide reasonably accurate estimates of

coverage error for certain demographic, socioeconomic categories at the state level and for the total population at certain substate area levels (large cities, SMSA's, some revenue sharing areas, etc.).

The emphasis in conducting the 1950 and 1960 postenumeration surveys was on obtaining data of good quality. Highly qualified staff were hired, given extensive training, and a considerable amount of time was devoted to seeing that procedures were properly conducted. The effect was to reduce errors due to poor enumerators and carelessly implemented procedures; however, the correlation biases arising from the tendencies of certain segments of the population not to be enumerated were largely unaffected (in fact, they may have been increased).

The emphasis in the 1980 program will be on independence from the census, in addition to quality. The 1980 program will utilize "independent" administrative files for purposes of improving the estimates of coverage error. To the extent that a satisfactory match between the administrative files and the census and CPS can be achieved without impairing independence of the sample data. we should be able to obtain more accurate estimates of coverage error

than were obtained in 1950 and 1960. The feasibility of using these administrative files is being investigated in a study currently underway. Data were collected from the persons in the February 1978 Current Population Survey (CPS) in order to facilitate a match with administrative files. Dual system estimates of the true total population will be made as of February 1978 and compared with estimates based on births, deaths, and previous censuses. If the two estimates of total population are reasonably close and the processing problems of administrative file matching are surmountable, administrative files will be used to adjust the CPS estimates of coverage error in the 1980 census.

The procedure for doing the August, 1980 CPS follows:

A listing is made of all Persons currently residing in the sample housing units together with all persons who died in there households subsequent to the census. A determination is made where each listed person was living at the time of the census. These addresses are then searched in census records to see if the sample persons were enumerated (Procedure B).

Since this procedure is only concerned with obtaining a roster of persons at the current address, we would expect this procedure

to yield a more complete listing and a better estimate of undercoverage than was feasible under the procedure used in the 1950 & 1960 evaluation surveys (A listing is made of all persons who resided at the sample housing unit at the time of the Census. The census records for the sample addresses are then searched to see if the persons were enumerated.)

In addition to estimating a gross omission rate from the CPS, we also plan to estimate erroneous enumerations in the census; therefore, the purpose of the Undercount estimation program will be to estimate a net coverage error, gross omissions minus erroneous enumerations.

A person is "correctly enumerated" if he was enumerated in the census at the address reported by the CPS as the census date residence. A person is "missed" if he was not enumerated at the census date residence that was reported in the CPS. An enumeration is considered to be "erroneous" if the CPS reports that the person was not living at the location where the census recorded him. For example, the CPS could report that no such person exists, or that the person was born after the census. died before the census or was living elsewhere on day. Also a person was erroneously enumerated

if he/she was enumerated more than once.

A separate sample of approximately 100,000 households will be selected from census enumerations to determine if they were erroneously enumerated.

3. Matching Techniques

One of the most difficult operations to design and implement is the development of matching techniques that involves:

1. matching of CPS housing unit and person records to census enumerated housing units and persons.
2. matching of CPS and census enumerated housing unit and person records to "administrative" file records.

These matching operations are different in that the

former involves a searching operation in a file arranged by address, whereas the latter involves searching files arranged on some other basis (in the case of the IRS and Medicare: files the search is on the basis of a social security number). Therefore, our research effort has taken different paths in determining optimum procedures for these two operations.

a. Matching of survey housing unit and person records to census records

The matching operations conducted for the Oakland, Richmond and Colorado postenumeration surveys* were clerical in nature with explicitly written matching rules. The Oakland PES was our first attempt to create a set of matching rules; since they were changed a number of times during the experiment, a definitive set of rules does not exist for Oakland. Based on our Oakland experience a set of explicit rules for persons was devised for Richmond and Colorado. The basic matching operation consisted of the following:

1. Coding the PES addresses to tract, ED, block, serial number, and form type. This information is needed to

locate initially the address in the census address register which then guides us to the corresponding census questionnaire. Maps with corresponding map-spotted units were used when searching for geocoding census addresses. Also the block header record that identifies the ED and block for a given street name and house number proved to be very useful when searching for census addresses. Telephone and city directories were used to a lesser extent in the searching operation.

2. Matching PES listed housing units against the census address register in order to obtain an estimate of census housing unit coverage.

3. Transcribing information from the PES interview forms to a special form to be used to control and facilitate the person matching.

4. Matching persons on the PES interview form to persons on the census questionnaires. Name, relationship, sex, age, date of birth, and race were used its matching variables for Richmond and Colorado.

5. For the Oakland PES, all Procedure B nonmatches, and

possible match cases were followed up to see if additional information could be obtained to determine match status for the "possible" match cases or to obtain additional address information for the nonmatch cases.

6. Lastly, a final matching operation to census questionnaires was conducted to determine final match status.

The following are general observations based upon our experience with the matching operations:

1. Follow-up (or reconciliation) will involve only cases for which additional CPS information is needed to determine match status. If the additional information cannot be obtained, the will be included as part of a noninterview adjustment and a search for a corresponding census record will not occur.

2. Matching in movers has been a difficult task.

Indications are that we were unable to locate a significant number of reported census day addresses (addresses other than the PES address); also, many addresses that were located were done so only with a great deal of difficulty. This experience was

especially noted in predominantly rural areas.

b. Matching of CPS and census enumerated housing unit and person records to "administrative" file records

Certain groups of persons are particularly likely to be missed by both the CPS and the census; examples are: black males, males in urban "ghetto" areas, low income adult males and migrants. Two administrative files are being used to provide alternative estimates to the CPS Census match coverage estimates for these groups. These files are the Internal Revenue Service (IRS) tax return file for persons of ages 18 to 64 and the Medicare file for persons of ages over 64.

The methodology to be used in forming a "triplesystem" census coverage estimate will consist of matching CPS records and a sample from census enumerations to the IRS and Medicare files. A brief description of dual-system estimation is explained later in this presentation. Matching will be done on the basis of a reported social security number (SSN). The Social Security Administration's alphadata and Summary Earnings Record File will be used to obtain SSN's for certain census and CPS records, and to validate reported

CPS numbers. This is discussed more fully in Section IV.

4. Administrative Matching

A possible improvement to using the CPS to estimate net Undercoverage in the census by a match to census records (dual-system estimation) is to additionally match to administrative records to form triple-system estimates. The two sources planned for use in 1980 are the tax returns filed in 1980 for 1979 fiscal year and the Medicare file of all Medicare records for the year 1980. There are several problems with using these files, the major one

*Special postcensal surveys (PES's) were conducted for the Oakland, Richmond, and Colorado census pretests for the purpose of estimating the census undercount.

being the size of the files; the IRS tax file alone contains about 85 million records, stored on 131 data tapes in SSN order.

Names and addresses are given to the Bureau exactly as they are listed on the tax return, meaning the address could be the address of the tax filer's bank, lawyer, or whoever prepares his tax return, or a family member. The Medicare file problems are similar, but on a smaller scale. Thus information may be reduced for confirming or negating matches.

To match to either of these files, it is necessary to have a SSN for the record to be matched. Note that this is true for records matched to the IRS or Medicare files, but not necessary if matching is done from either file to the census. The distinction will be clearer in a moment. The reason for needing the SSN is twofold:

1. Since the files are in order by SSN, it is most cost effective to search the files using that indicator. Matching to these files using names or other variables

would be prohibitively expensive.

2. The SSN is nearly a unique identifier. While one person may have several SSN's, possessing more than one SSN is a relatively rare event, and on the IRS files each SSN should belong to only one individual. However, identification using a person's name and matching in either direction can have problems when the individual possesses a common name (e.g., Robert Smith).

Unfortunately, for these purposes, SSN is not collected in the census, even on a sample basis. However, we plan to collect this information as part of the census erroneous enumeration survey, which is a sample taken from the census.

Matching can go in the other direction, too. A sample of cases with name and address can be drawn from the IRS and Medicare files and matched back to the census, in much the same way the CPS is matched to the census. However, problems with matching in this direction arise due to the need for a timely state sample; special arrangements would have to be made with IRS to draw a state sample while they are receiving return forms. This is necessary because the final IRS tax return file with names and addresses is not

available to the Census Bureau until approximately a year after the receipt of the forms. Also we have had some indications that special problems in matching could occur due to the nature of the Address that is filed with IRS, e.g., children who have moved away from home very often still file their parents, address as their residence.

It is also anticipated that a follow-up operation would be necessary because of the portion of the sample from IRS which would list an address used for tax return purposes which was not the residence as of census day. This could introduce a substantial bias into the dual-system or triple-system estimates by causing a low matching rate at the person's residence.

A supplement was administered as part of the February, 1978 CPS, collecting information necessary to matching the sample into the IRS tax return file for fiscal 1977. Dual-system estimates were developed from this matching project and are presently being compared to demographic estimates for 1978. This project should give us an indication both of the problems to be encountered in matching in this direction and will also tell us, by comparing the dual-system estimate to demographic estimates, whether the assumption of independence of sources in the dual-system estimate

holds.

5. Research Conducted for Proposed Match Study

a. February 1978 CPS/IRS match study

A special supplement was administered as part of the February 1978 CPS, collecting information necessary to validate and obtain SSN's at Social Security Administration. SSN's were then matched into the IRS tax return file for 1977. Dual-system estimates of the total population will be developed and compared with demographic estimates. This project should give us an indication of the problems encountered in matching, as well as whether the use of the IRS file for estimation purposes will lower the "correlation bias."

There are two separate operations that are involved in this match study. An operation at the Social Security Administration that involves validating and obtaining reported SSN's and a matching operation at the Census Bureau involving a SSN match of CPS records to the IRS files.

1. Social Security Validation Study--Social Security numbers were reported for about 80 percent of the eligible persons in the February 1978 CPS. Of the remainder, some SSN's were unknown to the respondent and could not be obtained by means of follow-up ; some respondents refused to report SSN; some persons reported that they did not have a SSN; and some SSN fields on the questionnaire were left blank without explanation.

Initially the CPS records with a reported SSN were validated at SSA by matching against the Summary Earnings Record file (SER).The validation was accomplished by comparing the first six letters of the surname, month and year of birth, sex, and race (the only comparable data in both the CPS and SSA files). The CPS records on which all comparable characteristics agreed with the SSA data, records with varying degrees of disagreement, and those records with reported SSN's that did not exist in the

SSA system were compiled for the Census Bureau. A further validation of records with varying levels of disagreement and CPS records that could not be located in the SSA numeric file was made by manually matching a sample of these records with a SSA alphabetic file. In order to test this procedure a test sample of 1000 CPS records with a validated SSN was also run simultaneously through the process with the valid SSN removed. Clerks were used to find these CPS enumerated persons, by name and date of birth, by searching in a microfilm file of all applications for SSN's. The CPS records included the following information that could be used in the searching operation.

- Person's full name and its corresponding soundex code
- Up to two previous or alternative names (maiden name, former married name, name before adoption, etc.)
- Date of birth: (month--day-century-year)
- Sex and race
- Mother's maiden name
- Father's name

- Place of birth (city or county, State or foreign country)

This information was included on a match form that included room for corresponding Social Security Administration data. An evaluation will be done to determine the extent of the use of the above information in determining match status.

The microfilm file-of SSN applications included the following information for each person:

- Soundex, code of the last name--(The soundex code is a device for grouping together spelling variants of the same name, and names that are spelled differently but sound alike and could easily be confused by an interviewer.)
- Last, first, middle name
- Date of birth: (month-year)
- Sex and race
- Social security number
- Mother's maiden name
- Father's name
- Place of birth

Records in the microfile file were arranged:

- By soundex code of the last name

- Within soundex group by first name and middle name (or middle initial)
- Within name group, by date of birth Confidentiality of all census forms was maintained by having the matching done by Census Bureau employees and having the study directed at the Social Security Administration by professional personnel who are census agents.

2. Match of CPS SSN's to IRS Tax Return File After the work done at Social Security Administration to validate and obtain SSN's. the CPS records were returned to the Census Bureau accompanied by a SSN. At the present time it appears that we will not be able to obtain valid SSN's for approximately 10 percent of appropriate CPS records (adults who could report on the IRS file). Since incorrect SSN's could still remain on this file, an additional validation study of SSN matching will be done; this will involve using name and address information that is available on both the CPS and IRS files, to determine the proportion of cases incorrectly matched by SSN. This is the first and only use of address information in the matching.

In order to obtain dual-system estimates, a tabulation of age, race and sex totals in the IRS file has to be prepared. This is being done on a 20 percent basis.

b. IRS-Census match study (involving Richmond, Va. and Southwest Colorado dress rehearsal censuses)

Approximately 1,000 tax returns were sampled from the IRS file for Richmond, Va. and approximately 1,300 sample cases from southwest Colorado. These were then matched to census records for these two areas. The purpose of the test was to determine if a match in that direction was feasible. Since the match is on the basis of name and address (no SSN is available for census records), we were especially concerned that IRS tax file addresses could result in a large nonmatch rate, resulting in a need for extensive field follow-up work. These results are now being evaluated. Preliminary indications are that this approach may be feasible and, in fact, more extensive tests. possibly on a national basis, could be warranted.

6. Estimation

The primary purpose of the estimation procedure is to provide estimates of the net Undercount for states (including the District of Columbia), and selected substate areas. A primary goal of the coverage evaluation program is to provide a methodology for determining corrected population counts at the state and substate area level. Since we cannot afford a survey to accomplish this objective at the local area level, we are developing a program that could be utilized in developing synthetic estimates at this level. Broadly speaking, this will involve two CPS samples that collectively will provide reliable estimates of the corrected population of specified minorities at the national level. The first estimates that could be formed after the census is concluded, would be dual-system estimates of the total corrected population for each state and for certain large SMSA's and cities. To obtain these estimates, the CPS will be matched back to the census, with the match

Table VI.3.1 Forming a Dual-System Estimate for one of the 61 Divisions

	Census		
Census Population Survey	In	Out	Total
In	M'	--	N'.P
Out	--	--	--
Total	N'.C	--	N'.T

where $N.T = \frac{N'.P * N'.C}{M'}$ is the dual-system estimate of the total corrected population for one of the 61 divisions.

N'.P is the estimate from the CPS of the total population;

M' is the estimate from the CPS of the number of persons enumerated in both the CPS and the census. adjustment is made for CPS nonresponse cases and for CPS insufficient information for matching cases:

N'.C is the total population count obtained in the census. minus the estimate of erroneous enumerations and of the total number of imputations made.

status ascertained for each person in the household. The CPS sample is being drawn as a state sample with supplementation of the largest SMSA's and cities. Each person or household in the sample will ultimately be classified as correctly enumerated, omitted, or erroneously enumerated. The sample estimates of the proportion of matches and of erroneous enumerations will be used in the dual system estimate to obtain the total corrected population in each of the states and designated SMSA's and cities.

The dual-system estimate is basically that used in capture-recapture methodology to provide population counts of migratory

animals, birds, and fish. Of necessity, one or two modifications have been introduced to allow for the vagaries of survey data. The estimate is formed as shown in Table VI.3. 1.

The only assumption required in this model is that the two sources be independent. If independence holds, then N'.T is the maximum likelihood estimate; N'.T is the final estimate of the total corrected population. It already allows for processing errors. census refusals and other cases which could not be matched since the cases are represented in N'.P but not in M'. To estimate the completeness of the census count or to estimate the census Undercoverage, we must add the imputations and erroneous enumerations back to N'.C. That is

$$PC = \frac{N.C}{N'.T} = \text{estimated completeness of census enumeration}$$

where $N.C = N'.C + E'.C + I.C =$ actual census count including erroneous

enumeration (E'.C) and

imputations (I.C)

$$\text{also } w.C = \frac{N'.C}{N'.T} = \frac{M'}{N'.P}$$

= proportion matched estimated completeness of the actual field enumeration, excluding erroneous enumerations and before any imputations.

Imputations and erroneous enumerations have to be excluded in estimating N.T because none of the imputations or erroneous enumerations will be matched and thus will not be included in M'.

Also using the above notation

$O.C = N'.T - N'.C$ is the number of persons not counted in the census.

$O.C = N'.T - N.C = O'.C - E'.C - I.C$ is the difference

between the total

corrected

population and

the census count.

$$q.C = I - p.C = \frac{O'.C}{N'.T}$$ is the net Undercoverage rate.

$$\text{and } r.c = I - w.c = \frac{O'.C}{N'.T}$$ is the gross undercoverage rate.

These procedures can be found in Marks, Seltzer and Krotki who also develop a three system estimator (2).

Following the work of Deming and Chandrasekaran (3), the dual-system estimate is formed for demographic subgroups within the region for which the estimate is being formed. These estimates are made for the smallest mutually exclusive demographic categories

(e.g., young black males), and added across categories to obtain the estimate for the region. This is done to reduce both the variance and the bias of the estimate.

These estimates would be revised as more information about the undercount becomes available from administrative record matching. Matching will be done using administrative records, and separate estimates of the undercount can be formed from a Census/IRS match and from a Census/Medicare match. These would be compared to the Census/CPS estimate and an adjusted estimate prepared. Demographic estimates for the U.S. as a whole will also be available. The state estimates obtained from matching can be adjusted to these national totals. As mentioned previously, there art timing problems in obtaining estimates from matching to administrative records, which lead to these estimates being produced later than the CPS estimates; hence the need for revisions.

A more complex estimator can be formed which involves a good deal more work. The concept of the dual-

system estimate can be expanded to comprise an n-system estimate, where now three sources are used in the matching process: the census, CPS. and a combination of Medicare records and the IRS tax return file. Matching problems faced in the dual-system estimate increase threefold because of the number of relations possible. Offsetting the increased matching problems, however, gains are made in both reduced variance and reduced bias when employing three systems. This is illustrated in work by Woltman and Smith (4) and Wittes (5).

7. Anticipated Cost and Timing of Administrative Record Match Study

Results of the two IRS studies should be available by August 1980. The costs of the studies are approximately:

A. IRS-Richmond/Colorado Match Study-Processing of 3,000.

records $13,000

B. CPS-IRS Match:

1. CPS Supplement involving 97,000 persons. Data collection-preparation $95,000

2. 'Computer matching to SSA numeric file 78,000 person-records $10,000

3. SSA Soundex Lookup involving 12,000 person records $6,500

4. Keying of SSA Lookup Records involving 5,700 person records. $500

5. Computer matching to the IRS file (involve's two passes of the IRS file, matching a total of 82,000 person records from CPS) $125,000*

6 Tabulations. $15,000

7. Other (salaries. etc.) $50,000

*This is a partial cost because the project was shared with other independent studies.

For more information on the 1980 Census Coverage Evaluation,

contact:

David Bateman

Statistical Methods Division

Bureau of the Census

Washington, DC 20233

8. References

(1) U.S. Bureau of Census. Census of Population and Housing: 1970. Evaluation and Research Program. PHC(E)-4. Estimates of Coverage of Population by Sex. Race. and Age: Demographic AnalYsis, 1973.

(2) Marks, Eli S., William Selmer, and Karol J. Krotki. Population Growth Estimation. A Handbook of Vital Statistics Measurement. New York: The Population Council, 1974.

(3) Chandrasekaran, C., and W. E. Deming. On a Method of Estimating Birth and Death Rates and the Extent of Registration. Journal of American Statistical Association No. 245 (March): 101-15, 1949.

(4) Woltman, Henry and William Smith. An internal Census Bureau Memorandum, Preliminary Finding on Dual vs. Triple System Estimation. June 4, 1979.

(5) Wittes, Janet T. Applications of a Multinomial Capture-Recapture Model to Epidemiological Data. Journal of the American Statistical Association, Vol. 69, p. 93-97, March 1974.

E. Case Study 4: Record Linkage in the Nonhousehold Sources Program

1. Introduction

The Nonhousehold Sources Program is a large-scale record check developed at the Bureau of the Census. The record check process is to match names and address records developed independently from the census to names and addresses collected in the census in order to identify persons who may have been missed in the census enumeration. The program will be carried out as an intrinsic part

of 1980 Decennial Census procedures in selected areas of the country. The basic purpose of the Nonhousehold Sources Program is to reduce within-household undercoverage and, in particular, to concentrate efforts on minority populations which are most likely to be undercounted.

The major steps in the Nonhousehold Source Program are:

(1) identification of the target geographic universe; (2) procurement of appropriate records, collected independently of the census, which specify names. addresses. and minimal demographic characteristics of in-scope persons, (3) precensal processing of the record lists to screen on geography and other characteristics of interest, and to prepare materials for matching; (4) a clerical match of the nonhousehold source records to census listings after completion of the first phase of census enumeration; and (5), follow-up of nonmatches to determine enumeration status whenever possible. In this last step. if it is determined that a given person had not been enumerated. he/she is added to the census questionnaire for the appropriate housing unit. As a further coverage improvement. the roster of persons reported for that housing unit is verified to add any other persons in the household

who we c missed in the initial phase of enumeration.

The Nonhousehold Sources Program is only one of several coverage improvement operations planned for the

Decennial Census. During the developmental phase for the 1980 census, several other coverage improvement programs have been initiated, expanded, and/or improved. The target population of many of these other coverage improvements overlaps with that of the Nonhousehold Sources Record Check; that is, a person missed in the early phase of enumeration may be added to the count from any of a number of coverage checks. Therefore, it has been determined that the Nonhousehold Sources Program will have the greatest payoff, in terms of coverage improvement, by checking a great number of cases under relatively liberal criteria than by checking for fewer cases

under strict or conservative rules.

The Nonhousehold Sources Program has been extensively pretested in planning for the 1980 census. Procedures evolved beginning with the Travis County, Texas census test in April 1976; going to the Camden, New Jersey pretest in September 1976; continuing with the Oakland, California pretest in April 1977; and finally in the Dress Rehearsals (Richmond, Virginia, April 1978; and Lower Manhattan, New York, September 1978)..1 As of this writing, only the Travis and Camden results have been analyzed in detail. In Travis, 7.5 percent of the persons from the record lists could have been added to the census, but a mechanism to actually change the counts was not yet developed. The equivalent number for Camden was 6.3 percent of the lists. In Camden, the missed persons were actually added to the census counts. The "yield" of the Nonhousehold Sources Program was even higher, as a number of persons were added as a result of the roster check at follow-up households. In Travis, an additional 3.3 percent, on the base of the number of persons record checked, was added through the roster check giving a total of "yield" of 10.8 percent of the list. In &;Zen, an additional 2.5 percent of the list was added, giving a total "yield" of 8.8 percent. Results for the other tests will be

forthcoming.

Based on data available to date, the results of the program with respect to coverage improvement were sufficiently encouraging so as to lead to the inclusion of the program in the 1980 census. In 1980, plans are to record check 7,000,000 names and addresses of persons in urban areas of minority concentration. The independent record sources will be lists of holders of drivers licenses, supplied by the various States; and lists of persons from selected countries of origin and registering as resident aliens in January 1979, supplied by the Immigration and Naturalization Service. Drivers license lists are a desirable nonhousehold source bemuse: (1) they are public records and therefore are fairly readily obtained from the States.2; (2) they are universally computerized, and thus facilitate mass processing; and (3) name and address information is relatively recent--only licenses with reported addresses less than two years old are used. In addition to drivers licenses, a smaller number of cases will come from the registered alien lists, which have the same advantages as license lists. The INS lists were first used in the Oakland, California census pretest; although definitive results are not yet available from

Oakland, preliminary counts indicate the yield from the INS list was similar to that from drivers licenses. The INS lists contain not only the same name, address, and demographic data as the license lists, but also supply "country of origin" so that appropriate race/ethnicity screening can be done.

2. Results from the Travis County, Texas and Camden, New Jersey Pretests

The reminder of this report will concentrate on the matching phase of the Nonhousehold Sources record check, as studied using the results of the Travis County. Texas and Camden, New Jersey census pretests.

For the Travis County pretest, a total of 3,002 names and addresses went through the entire record check procedure. Of these, 2,342 cases were from drivers licenses. For cost purposes, the Travis driven license cases were confined to males, aged 17-35, in two Zip code areas of Austin City identified as having high minority populations. The additional 660 names and addresses were supplied by local community organizations. These encompassed both

sexes, a larger age range, and more geography. The names and addresses were transcribed to the control section of an office worksheet, to be used later in geocoding, matching and recording follow-up results.

In the Travis local census office, the addresses were assigned census geographic codes (geocodes). this was done successfully for 2,910 of the original 3,002 records. Once a geocode was assigned, the worksheets were matched to the master address listings for the a geography. A serial number identifying the address was located and the census questionnaire for that serial number was obtained. The name from the record source was matched to the household roster on the questionnaire to determine if the person had already definitely been enumerated, or if further follow-up efforts were necessary.

The address and name matching rules used in both Travis and Camden can be found in the Appendix (section

---

.1Actually. the first attempt at a large-scale Nonhousehold Sources Record Check was in the context of a special census

conducted in Pima County. Arizona. in 1975. However, because the procedures used varied considerably from those in the pretests and the census, the results of that check will not be discussed here.

.2In developing the 1980 program, some States have cited Privacy Act restrictions in denying records to the Bureau. However, in further discussions, this limitation is found not to apply since the records are treated in accordance with Title 13 when in the Census Bureau's possession.

7). For Travis, in the match between the geocoded addresses and census records, 2,719 or 93.4% of the initial 2,910 geocoded addresses were successfully matched; 86 or 3% were called possible matches; and 105 or 3.6% were nonmatches. The possible matches were eligible to go through the name match and further follow-up efforts; nothing further could feasibly be done with no matches.

In the match between the 2.805 names on address matched or possibly matched records and the census questionnaire rosters, 1,378, or 49%, were classified as name matches. 159, or 6% were possible matches; and 1,268, or 45%, were nonmatches. The possible name matches and nonmatches were sent for telephone and, if neces-

sary, personal visit follow-up to obtain further information. As a result of these follow-ups, 207 of the 1,427 unmatched persons were determined definitely to have been missed; 154 were out-of-scope (deceased, moved from test area, etc.); and, for 1,046 persons, enumeration status could not be determined.

For the Camden Nonhousehold Sources Program, as in all later efforts including the census, the geocoding of addresses from the drivers license lists was done by computer (the 275 in-scope names on local lists were hand geocoded). In order to answer questions regarding yield rates for different demographic groups, the nonhousehold sources sample was allocated such that all adult persons, male and female, were represented, although the emphasis was still on younger males. After geocoding the license lists supplied by New Jersey, a total of 19,840 records remained, from which a stratified sample of cases was selected. A small unstratified sample of addresses not geocoded by computer was also included. In all, a total of 6,099 cases were processed through the Nonhousehold Sources Program in Camden. The names, addresses, and geocodes were printed on record search forms and sent to the local office to be matched to the census roles.

As in Travis, the clerical match was performed in two steps: first on address, and where that was successful, on name. For address matching, note that the categories of address matching had been expanded. This change came about because of two problem situations noted in Travis. When the record address was matched to a unit which was "vacant" or "deleted" on the census list, there was no place on the Travis form to indicate the situation. In such cases there is no reason to go to the census questionnaire to attempt a name match. Therefore, it was decided that such cases would be noted and set aside, the assumption is made that other census operations would add the household members to the address, if appropriate. The second problem arose when the basic address on the nonhousehold source record matched to a basic address for a multiunit structure in the census, but the independent source gave no apartment designation. When this happened, it was not possible to readily identify the serial number of the appropriate census questionnaire for the name match. In Travis, this was handled by searching all the questionnaires for the basic address; in Camden, the category "multi-unit structure" was added to allow the matchers to indicate when this was done.

The result of the Camden nonhousehold sources

address match showed 5,763, or 94.5 percent, of the 6,099 cases were matches; 360 of these matched to a vacant or deleted unit. There were 18 (0.3% of 6,099) possible matches and 224 (3.7% of 6,099) addresses that matched to multi-units with no apartment designation. Only 94 cases were non-matches. Those 5,645 cases which were not classified as address nonmatches of matched to vacant or deleted units went on to the name match.

The Camden name match categories were also expanded with the addition of the "Unable to Locate Questionnaire" classification. The nonmatches were postcensally classified to separate the cases where the name could not be matched because the household was a refusal in the enumeration, from cases where the person was just not matched to an existing roster. The results of the Camden name

match were as follows:

It can be seen that 2,574, or 45.6 percent of the names matched initially a result almost identical to Travis. In Camden, however, it was possible to examine the match rates by the three demographic groups shown above as represented in drivers licenses. It can be seen that females matched at a higher rate than males, and that males 25-44 matched at a much lower rate than the other males in the sample.

Postcensally, a more thorough review of the matching operation was carried out. Of the 227 "Possible Matches' " 154, or 67.8 percent of the person were eventually verified enumerated, 61, or 26.9 percent were undetermined, and 12 persons (5.3 percent) were added to the census. This distribution supplies much of the argument for the eventual elimination of the "Possible Match" category. Another postcensal study looked at the consistency of drivers license records with census data, and the accuracy the initial name matching operation. The name matching criteria used in the initial office matching operation did not require an examination of the answers to the census age or sex questions to

establish a "Match." To evaluate the consistency between the age and sex of nonhousehold sources cases that were classified as a harm "Match" and the corresponding age reported to the Census Bureau, the census questionnaires for 2,338 cases classified as a name "Match" in Camden were reexamined. It was first determined if, in fact, the cases were a name "Match" according to the Camden matching criteria. Of the 2,338 cases studied, 22 (0.9 percent) had erroneously been classified as matches in Camden. There were 42 cases which had erroneously been called "Nonmatches," for a gross error or total of 64 (2.7 percent of the total number of cases studied plus erroneous nonmatches). This result indicated that, even though the matching clerks had minimal training and supervision, the matching rules were applied relatively well.

A comparison was the made for "sex" as reported from drivers licenses and the Census for the 2,316 names correctly matched in Camden. Of these, sex differed on the two sources in 15 cases. This error could have come from misreporting or misallocation on either source.

Age was then compared on the two sources for the 2,301 cases found to be name matches of the same sex. The following table

displays the result of this match:

The above table presents a cross tabulation of age reported on the drivers licenses and that on the Census questionnaire. Along with row and column distributions, diagonal totals are presented and summarized. The amount of agreement noted is evidence of the quality of age reporting on drivers licenses, as well as the accuracy of the matching operation. Within the age ranges tabulated, 93 percent of the cases fall on the diagonal and an additional 3 percent fall within one cell. For the off-diagonal cases, we suspect a large number of there arise because a parent's name was matched to a child's, or vice versa. For program purposes, these imperfections are acceptable. It would not be worth additional time, training, and follow-up effort to resolve such discrepancies. The result of not reconciling these differences is that a minute amount of coverage improvement may not be realized, but the cost of reconciliation would be prohibitive.

3. Plans for the 1980 Census Nonhousehold Sources Program

In the Oakland pretest and the Dress Rehearsals, further modifications were introduced into the nonhousehold sources matching records. The procedures and forms for the 1980 Nonhousehold Sources Program have evolved on the basis of these pretest and dress rehearsal experiences. The section for recording match results has been expanded to cover all relevant situations, and procedures for how to handle each case appear appropriately (see Section B):

The basic matching instructions have been modified somewhat from the pretests. The "Possible Match" category has been dropped. This was done because relatively few cases were categorized this way in the pretests; more importantly, however, "Possible Matches" would be treated, at each point, just like matches. Also, to handle the problem of no apartment designation appearing on the record source when the census shows a multi-unit structure, a distinction is made which depends on the size of the structure. For basic addresses with ten or more units, nothing further is done. For those with fewer than ten units, an attempt is made to

identify the correct unit by matching the surname to the census. This is done to keep the operation workable and to keep the matching clerks

honest. However, if the person is not found, no follow-up is possible without a specific unit to call or visit.

4. Summary and Future Considerations

In summary, it is felt that the match of administrative ("Nonhousehold") records and the census is sufficiently accurate to meet the aim of cost-effective coverage improvement. Perfection in procedures and accuracy in the independent record source have been shown unnecessary in generating a highly acceptable yield from the processes involved. Perhaps the most disturbing aspect of the results of the program is the large number of "undetermined status" cases which have consistently arisen in pretests. Given that the matching procedure itself is accurate to the degree it was in Camden, the fact that enumeration status is never determined for at

least one-fifth of the records checked must be a function of procedures other than the clerical match. It is probably a function of incorrect addresses reported to the Department of Motor Vehicles; the inability to conduct a follow-up interview in the census; the mobility of the population; an unwillingness of the target person to be interviewed; and other factors. The degree to which each of these factors contribute to the "undetermined's" is a subject for further research.

One last word regarding the choice of administrative lists to use in the Nonhousehold Sources Program might be appropriate. The previously discussed requirements--currency, availability, computerization, presence of minimal data--are met by drivers license and INS records. The experience of using locally-supplied lists in

Travis and Camden showed there to be costly to use on a large scale basis and, more importantly, less effective in terms of percentage yield than drivers licenses. It has often been suggested that some form of Public Assistance lists be used, as these might be fruitful to enumerate the types of persons likely to be missed. In fact, for the final Dress Rehearsal in Lower Manhattan, a welfare file (comprised of recipients of AFDC, General Public Assistance, and Medicare) supplied by the city was used. The results of its use, when available, will indicate whether this list may give a higher yield rate. However, such lists may be very difficult to procure, particularly when they are controlled at local rather than State levels. They are also protected to a great extent by privacy laws and provisions; for instance, the Department of Agriculture has denied access to Food Stamp Roles. However, in spite of the decision to use drivers license and alien lists in the 1980 census, the issue of which administrative record sources to use is not closed. It is expected that efforts to improve the Nonhousehold Sources Program will continue into and beyond the 1980 census, and the investigation of other list sources will undoubtedly be a part

of them.

5. Sources of Further Information

Further information regarding the Nonhousehold Sources Program is available as methodological research documentation at the Bureau of the Census. After the 1970 census, a small scale evaluation study of the use of drivers licenses as an administrative record source was performed in Washington, D.C. The results of this study are described in the 1970 Census Preliminary Results Memoranda Series. [1]

Original interest in this program, and preliminary recommendations for implementation, may be found in the memorandum series of the Task Force on Coverage Improvement Procedures, active after the 1970 census. [2] Credit for the original tabulation and analysis of results for the Travis and Camden Nonhousehold Sources Program is given to John Thompson, Statistical Methods Division. Further discussion of the Travis and Camden programs can be found in three memoranda by that author. [3], [4], [5].

A comprehensive summary of these results can be found in a

paper entitled, "The Nonhousehold Sources Coverage Improvement Program, " presented by Thompson at the American Statistical Association Annual Meetings, August 1978. [6]

The extensive computer programming efforts for the Nonhousehold Sources Program have been carried out under the direction of Roger Lepage, Decennial Census Division. Information on processing the drivers license and INS lists, including the geocoding match, may be obtained from Lepage.

An overview of 1980 census coverage improvement efforts, including the Nonhousehold Sources Program, may be found in a paper, "Plans for Coverage Improvement in the 1980 Census," by Peter Bounpane and Clifton Jordan, presented at the American Statistical Association Annual Meetings. August 1978. [7]

For more information on the Nonhousehold Sources Program, contact:

Susan Miskura

Statistical Methods Division

Bureau of the Census

Washington, DC 20233

6. References

Copies of all documents cited below may be obtained from the Research Documentation Repository, Statistical Research Division, U.S. Bureau of the Census.

[1] Novoa, Ralph (1971), "Preliminary Evaluation Results Memorandum of the 1970 Census. No. 21. Subject: Listing Census Coverage through Drivers Licenses (E22-No. 3)," October 21, 1971.

[2] Marks, Eli S., Jones, Charles D., Cullimore. Stanley O., and Foster, Barbara (1974), "Memorandum for the Task Force on Coverage Improvement Procedures, Subject: Proposal for Use of Nonhousehold Sources for Coverage Improvement," October 18, 1974.

[3] Thompson, John H. (1977), " 1967 Census of Travis County Results Memorandum No. 34. Subject: Travis County Nonhousehold Sources Program. December 8, 1977.

[4] _____ (1977), "1967 Census of Camden. New Jersey Results Memorandum No. 15, Subject: Primary Results of the Camden

Nonhousehold Sources Coverage Improvement Program," October 28, 1977.

[5] _____ (1978), 1976 Census of Camden. New Jersey Results Memorandum No. 24. Subject: Additional Results of the Camden Nonhousehold Sources Coverage Improvement Program." October 25, 1978.

[6] _____ (1978), "The Nonhousehold Sources Coverage Improvement Program," 1978 Proceedings of the Social Statistics Section, American Statistical Association, 1978, 435-440.

[7] Bounpane, Peter A., and Jordan, Clifton (1978). "Plans for Coverage Improvement in the 1980 Census, " 1978 Proceedings of the Social Statistics Section. American Statistical Association. 1978. 12-20.

7. Appendix. Matching Instructions

1. Address Match Terms

A. An address is considered matched under the following conditions:

1. The identical street name, house number, apartment number (if any), State and Zip code appear in the register, or the house numbers are the same and the street names have only minor spelling variations. For example. "Freeman St." vs. "Freemen St."

2. The identical Post Office lockbox number, State and Zip code appear in the register.

B. An address is considered possibly matched under the following conditions:

1. The house numbers and street names appear to be the same, but the street types are different. For example, the word "Street" in one source and the word "Avenue" in the other source. This includes variations,between "Road," "Court," "Circle," etc., as well as a street type in one source but not in the other.

2. The house numbers and street names appear to be the same but the compass point is present on one source and absent in the other. For example. "301 Main St." vs. "301 N. Main St"' This DOES NOT include contradictory compass points such as "E. Oak" vs. "W. Oak".

3. The house number and street name were matched or possibly matched but the identical apartment number, letter, or location description is not found.

4 . The house numbers appear to be the same but digits may have been transposed in the register. For example, you are searching for the number "382" and do not find that number in the register, but find instead

"380 Elm Ct."

"328 Elm Ct."

"384 Elm Ct."

"386 Elm Ct."

Note that the sequence of listings provides evidence of transposition.

C. An address that is neither matched nor possibly matched is considered a nonmatch.

II. Name Match Terms

A. A name is considered matched when both a given and surname are shown in each source and one of the following conditions exist:

1. The names shown in both sources are identical.

2. The names are pronounced the same but are spelled differently. For example, "William A. Ralph" vs. "William A. Ralf".

3 . An abbreviated name is provided on one source and is noncontradictory to the name provided from the other source. For example, "Jim E. Johnson" vs. "James E. Johnson".

B. A name is considered possibly matched when one of the following conditions exists:

1. Only surname is given in one source and that surname is identical to the surname in the other source.

Slight differences may exist as long as they may be attributable to errors in spelling or handwriting.

2. Surname and one or more initials, but no given names appear in one source and that surname is identical to the surname in the other source and the initial(s) are noncontradictory. Slight differences may exist as long as they may be attributable to errors in spelling or handwriting.

C. A name is said to be a nonmatch if it is not one of the above.

F. Concluding Comments

The case studies presented illustrate the actual and potential benefits and difficulties involved in ca trying out studies using matching of administrative records to obtain statistical data. We will highlight some of the main issues raised by these studies.

1. The case study on the Linked Administrative Statistical Sample (LASS) project is intended to illustrate some of the main

concerns being addressed in order to determine the feasibility of developing integrated sa triples from several administrative record systems. In LASS, the main use of sampling from administrative records will be to create an improved database for industrial and occupational mortality research. There are at least three major issues which will have to be resolved before this objective is accomplished:

1. access restrictions and disclosure issues,

2. potential incompatibilities among the systems being linked, and ,

3. problems of data quality.

The suitability of an upgraded CWHS for studying industrial and occupational mortality depends, in part, on the results of

efforts to:

1. Add cause of death and other death certificate information to the CWHS. (It is not known yet if SSA information for decedents on name, social security number, race, sex, date of birth and date of death is sufficient for the States to attempt a search for the death certificate.)
2. Create detailed occupation codes from the occupation entry on individual tax returns and SSA industry information. (The usability of the occupational entry is being assessed given the lack of taxpayer instructions for reporting occupation.)
3. Upgrade the CWHS data on industry and place of work. (Data quality problems exist partly because of the voluntary nature of the SSA establishment reporting system. Other data quality problems are being encountered in the changeover to annual wage reporting.)

Access questions, though, are among the most important issues that have to be addressed before the Continuous Work History Sample can be used to its fullest potential for mortality research.There are at present many restrictions imposed on data access by laws

such as the Privacy and Freedom of Information Acts as well as the statutes and regulations of each of the participating agencies. Interagency data sharing is very limited, as a result. If Social Security is to proceed with the numerous activities planned for upgrading the Continuous Work History Sample, many confidentiality restrictions will have to be overcome. Legislative initiatives to resolve problems of making information available for statistical linkage (i.e., tax return data) and Presidential proposals aimed at providing government-wide legislation for protection of statistical and research data offer possible solutions.

2. The Use of Administrative Records in the Survey of Income and Program Participation describes the difficulties encountered in using administrative records as sampling frames in three experimental field activities prior to the ongoing SIPP. Three major problems have arisen in the SIPP development work. First, locating sample cases in the field has been more difficult than initially anticipated. The source, of this difficulty stems from several causes: (1) inadequate or inaccurate addresses, (2) recent moves by program participants, and (3) a minimum delay of several months from sample selection to interview date. Field procedures

have been adopted to help minimize the problem. These procedures seem to have improved the interviewers ability to locate the sample person; however, a further analysis of the procedures' impacts would be useful.

Unlike the first problem which became apparent in the survey field operations, the two remaining problems were first observed while investigating potential administrative record systems for the SIPP. Thus, the second problem concerns finding administrative record systems which are national in scope and relevant for the study of current policy issues. Few systems of interest for sampling maintain records at the national level. Systems available only at the State or local level would substantially increase the sampling and data access problems of the ongoing survey.

Finally, the third problem concerns the identification of sampling units with a known probability of selection. This arises when the survey unit does not coincide with the administrative data units. Some modification of the sampling frame is necessary to ensure a well-defined probability of selection.

In the developmental program, the main use of sampling from administrative records in the SIPP has been for validation studies

and response error analyses conducted by comparing survey data with administrative data. In the future, however, the main uses of administrative records systems will be for improving estimates for particular segments of the population through multiple frame weighting and/or for augmenting the survey data base with data which is difficult to collect, such as work history or earnings history data. Ultimately, data from administrative records may, be used to adjust individual survey data or to develop control totals for adjusting aggregate estimates of recipiency of particular income types and participation programs.

3. The case study on Use of IRS/SSA/HCFA Administrative Files for 1980 Census Coverage Evaluation serves to illustrate the difficulties of matching when different units are being linked and the identifiers differ. The CPS identifies households by address and includes SSN for household members; the 1980 Census also identifies households by address, but does not include the individual's SSN; the IRS/SSA/HCFA administrative record files list persons with the SSN as an identifier. Matching of the CPS records and the Census records is based on the geographic location of the household units; after a potential matched household was identified

based on the address, the characteristics used for matching individuals were name. relationship, sex, age, date of birth and race. For the match of CPS records with IRS/SSA/HCFA administrative records the main identifier used for matching individuals was the SSN; if a match could not be established on the basis of the SSN, then other identifiers were used, such as date of birth and last, first and middle name. The research study conducted using the February 1978 CPS match to IRS records showed that a valid SSN was

maintained for about 10 percent of the individuals' CPS records.

The main use of the 1980 coverage estimates is to estimate the undercount of the official Census estimates that are published in January 1, 1981. The estimates of 1980 coverage of population in the 1980 Census at the State, Substate level. and for selected

subgroups would not be available at the same time as the Decennial Population counts were released: when they become available they could be of crucial importance in the distribution of billions of dollar of Federal money based on population data (e.g. in programs such as General Revenue Sharing and various grant-in-aid programs). Another important use of the estimates could be in the program to develop intercensal population estimates.

4. Record Linkage in the Nonhousehold Sources Program was developed through a series of pretests for the 1980 Census of population. This program is a good illustration of the need to modify matching procedures over time after problem areas have been identified. These modifications will enhance the efficiency of this program in the 1980 Census.

To choose only administrative record lists with highest quality was an important decision in this program; the quality of the list influences to a great degree the proportion of matches which can be achieved. Locally supplied lists are of uncertain quality, difficult to use because they are not computerized. and overall were not as efficient as Drivers' License lists available from the States. In the final 1980 Census operation only cases.

with geocodable addresses went through the clerical name matching operation. About 8,000,000 names will be matched in this operation in areas selected for their high concentration of minority population. Because of the magnitude of this project. the list of households to undergo the clerical name matching was created by computer based address matching. Inaccuracy in name matching (i.e.. false nonmatches) leads to further field follow-up. which increases cost and time. However, because the match is to augment the Census rather than to make statistical estimates. degree of matching uncertainty is acceptable.

The Nonhousehold Sources Program is one of several coverage improvement program for the 1980 Census; therefore, it need not yield "perfect" results. It is only required to be a cost-effective operation to improve the coverage of the Census.

CHAPTER VII

Technical Problems in the Statistical Use of Administrative Records

Previous chapters have discussed the importance of administrative records for such uses as the generation of current statistics for small geographic areas. There are a number of technical problems which have been encountered with past and current uses that must be resolved if the statistical potential of administrative records is to be realized.

The technical problems are similar to those encountered in the use of census and survey records. With each administrative record set to be used, the statistician must ask: Who is reported? (is the appropriate population of persons. organizations, etc., fully covered in the record set?); What is reported? (Is it appropriate for the intended statistical use?); How is it reported and processed? (Is the information accurate?); When is it reported and processed? (is the information timely?). The unique aspect of

administrative record uses is that the questions are asked after the record collection has already taken place.

Administrative record sets are often not designed for statistical purposes. They may not cover the entire population of interest, they may contain administratively convenient concepts and definitions that are not appropriate to the statistical use, there may be lack of adequate control over the accuracy of key information, and it may be difficult to access the records in a timely fashion. Some of these problems are inherent in the nature of data processing in general (such as keypunching errors), but most can. with greater attention and planning, be resolved. in fad, a resolution of technical problems preventing effective use of administrative records for statistical purposes can, in many instances, improve administrative efficiency as well as produce better statistics. This is particularly true when technical problems impeding statistical use of records arise because of duplicative and inconsistent reporting requirements associated with different administrative programs dealing with overlapping populations. The political barriers to improved coordination among administrative programs, and between statistical and administrative programs, may not be easy to remove; but, in a number of areas,

improved coordination offers the potential for higher quality data, more efficient administration, and reduced reporting burdens for individuals and organizations.

This chapter will illustrate common technical problems that arise in making statistical use of administrative records by using the Social Security Administration's Continuous Work History Sample and related administrative record files as the principal examples. The CWHS has the advantage of making extensive use of administrative files containing both individual and organizational records; and it also has been used in a number of recent tests designed to assess the quality of administrative records for statistical applications. (See Chapter III for a detailed description of CWHS data programs.) The remainder of the chapter is organized into four main sections dealing with problems relating to: (1) incomplete coverage of administrative record sets; (2) lack of comparability among record sets; (3) reporting and processing errors: and (4) questions of data timing. A summary of problems and potential for improvement concludes the chapter.

A. Coverage

Information for nearly all employed persons is contained in one or more of the administrative record systems currently in existence, either because they have accrued income subject to taxation or are eligible for benefits under one or more Federal programs. There is not. however, a single record system containing information for the entire U.S. population. The statistician often must deal, therefore. with differences between the population of interest for the statistical purpose and the population covered by the record system. Such "coverage gaps" sometimes are difficult and costly to correct (although not nearly as costly as sample surveys to collect comparable information at detailed geographic levels.).

Most administrative record files contain information for specific groups such as persons receiving public assistance payments under a particular program. There are, however, at least three major record systems which cover large segments of the population: Internal Revenue Service records from income tax returns; Social Security

Administration records; and records collected by State agencies for unemployment insurance purposes. Each of these record systems has complex coverage limitations defined in terms of groups specifically excluded from mandatory participation in the administrative program.

Annual employee-employer CWHS files have been assembled principally from three major SSA administrative files--(1) a file of personal characteristics, which contains information on sex, race, and date of birth taken from applications for social security numbers; (2) an employer file which contains industry and geographic codes for employer reporting units, taken from applications for employer identification numbers and related supplemental informational forms; and (3) a wage item record file which contains

worker wage information taken from regular employer reports of individual wages subject to the social security payroll tax The personal characteristics file covers all individuals with social security numbers, the employer file covers all employers with employees subject to the social security payroll tax, and the wage item record file covers all individual wage and salary jobs subject to the payroll tax.

The personal characteristics file covers nearly all adult Americans, but information on earnings and location and industry of employment is available for individuals only for periods in which they work in social security-covered jobs. Nonworkers and those working only in noncovered jobs do not appear in the annual employee-employer CWHS file. The largest groups of noncovered jobs include Federal civil service and railroad jobs-which are excluded because of coverage by alternative pension plans-and jobs in those State and local government and nonprofit organizations that have not chosen to take the optional coverage available to such organizations. Most self-employed workers are covered by social security and a file of self-employment records can be merged with the employee-employer file.

The significance of the coverage limitations of the CWHS

depends on the desired applications of the data. Since the employee-employer CWHS provides information only for covered workers. contains only covered wage income. and provides no measures of family status or family income, it would be a seriously deficient data base for analyzing the overall economic welfare of particular demographic groups.

The CWHS files have been used to develop statistics on regional workforce characteristics and inter-regional migration. One of the results of the coverage exclusions is that persons who move between covered and uncovered employment appear as contracts to or dropouts from the workforce, thus overstating both of these categories and understating true migration. Another effect is that because of workforce composition, coverage will tend to vary from region to region. Workforce estimates for an area like Washington, D.C., with a larger concentration of noncovered Federal workers. are therefore deficient. Similarly, workforce and migration estimates for areas with high concentrations of noncovered State or local government employees are adversely affected.

One approach to resolving "coverage gaps" is to merge micro records from different administrative record systems. There have,

for example, been test efforts to merge Federal civil service employment records and Railroad Retirement Board records with the CWHS. These efforts were complicated, however, by noncomparabilities between the files in records relating to such important information as wages and salaries (the CWHS shows covered wages received, while civil service records indicate only grade level and rate of pay). Greater coordination of recordkeeping procedures among administrative agencies would facilitate data mergers. The Civil Service Commission, for example, is considering statistical applications in its design of a new information system and consequently may include payroll as well as personnel information in the records.

Matches of different administrative record sets for statistical purposes would help to overcome coverage problems and greatly improve the usefulness of data sets. As has been indicated in Chapter VI. however, both technical problems and problems of access can prom serious barriers to successful statistical projects to link records from programs under different administrative jurisdictions. The linkage barriers are particularly serious in such cases as State-administered welfare programs. where there may

be significant State-to-State variations in recordkeeping practices and access restrictions.

The most promising route to more complete employment coverage in the CWHS is related to the recent shift to a joint SSA-IRS program to collect a single annual report on individual wages in place of the current annual IRS report (form W-2) and four quarterly SSA reports. This coordinated record collection program ism Chapter V for a detailed discussion) is designed primarily to reduce the reporting and paperwork burden for employers and to improve administrative efficiency. but it potentially makes available for the CWHS a virtually complete set of annual reports for all wage and salary jobs.

In general, greater coordination among administrative agencies in record collection and processing should not only reduce paperwork burdens, but should also make administrative records more suitable and cam" to use for statistical purposes. The Commission on Federal Paperwork (1977), which initially recommended coordinated annual wage reporting to IRS and SSA. has, for example, also recommended greater interagency coordination of record collection for welfare benefit programs. If im-

plemented, their program (Single Application for Verification of Eligibility or SAVE) could potentially make it much easier, from a technical point of view, to merge welfare records with other records such as those in SSA and IRS files, in order to obtain a reasonably comprehensive statistical picture of individual and family income from administrative records. In order to insure that administrative recordkeeping changes reduce rather than increase noncomparabilities among record sets and do not add to other technical problems impeding statistical use of administrative records, however, there must be effective coordination between statistical and administrative agencies as well as coordination among administrative agencies. The next section discusses data comparability and quality problems arising from imperfectly coordinated programs for establishment reporting of employment and

payrolls in administrative and statistical systems.

B. Comparability

The statistician is often faced with the problem of adapting administrative record concepts and definitions to statistical needs. Not only do administrative concepts frequently differ from statistical concepts, but they can also differ among administrative record systems. One consequence of these differences is that measurement of the same phenomena (employment by industry, for example) will yield different results from the different administrative record systems. Reconciliation of the differences (and consequently an assessment of accuracy) is extremely difficult and complex. Another factor is that concepts are often interpreted and implemented differently by the various reporting entities in the same record system.

One of the primary uses of administrative records, as noted earlier, is the production of statistics for subnational areas. An important conceptual problem in the use of such statistics is that

some record systems measure economic activity at the individual's place of work whereas other systems measure activity at the individual's place of residence. Social security and unemployment insurance data, since they are based on employer reports, reflect place of work. Decennial census and IRS (1040) data generally reflect place of residence.

To illustrate the effects of data comparability problems that arise in using administrative records to develop employment estimates, Table VII.1 compares first quarter 1970 CWHS employment estimates from the 1970 census. Also compared for the Nation and New York State are first quarter CWHS employment estimates for 1971 and 1975 with employment estimates based on unemployment insurance records and employment estimates from the Census Bureau's County Business Patterns program.

The CWHS and decennial census estimates have a number of noncomparabilities. The census is by place of residence, whereas the CWHS is by place of work. The census estimates are based on questions regarding the person's employment during the week prior to the census while the CWHS counts persons with covered employment at any time during the first quarter. The census estimates include

self-employed persons while the CWHS estimates include only covered wage and salary workers.

The 1971 CBP data are derived primarily from social security records, and thus should cover essentially the same workers as the CWHS. There are, however, some important differences between the two series. The CWHS employment estimates have been tabulated from a 1 percent sample of records for individual workers and represents an estimate of workers in covered employment at any time during the first quarter of 1971, classified on the basis of the location and industry of their major job during the period. The CBP estimates, by contrast, represent a count of jobs filled during a single (mid-March) pay period derived from employer reports of aggregate employment submitted along with quarterly payments of social security payroll taxes. The CBP estimates moreover are based on regular SIC industry coding conventions and omit the government SIC category (because of incomplete coverage), while the CWHS estimates include government workers coded to nongovernment SIC categories whenever the government reporting unit is engaged in activities for which there are corresponding private SIC categories (e.g., school district employees would be coded into the educational services

industry).

An additional important difference between CWHS and CBP employment estimates arises because of Census Bureau supplementary efforts to obtain more complete and reliable industrial and geographic detail that can be derived from basic reports to SSA. Industrial and geographic breakdowns of employment by multiestablishment employers are supplied to SSA on a voluntary basis, and Census has long had its own program to collect geographic and industrial employment data from large multiestablishment employers that have not voluntarily supplied the information to SSA. The supplemental information supplied to Census is incorporated into CBP's estimates, but it does not contain individual worker records and cannot be incorporated directly into the CWHS. Therefore, the CWHS contains some geographic and industry distortion because of incomplete establishment reporting.

CBP data for the years 1974 and later are based on SSA data for single establishment employers only. Data for multiestablishment employers are obtained from the Census Bureau's Annual Organizational Survey which is used both for data collection purposes and to update the Standard Statistical Establishment List

(see Chapter V for a detailed discussion of the SSEL program).

The UI data in Table VII.1, like the CBP data, represent a count of total jobs held at reporting establishments during a single mid-March pay period. The establishment concept used in the UI system is generally consistent with that used by SSA (industry-county combinations), but it is an independent reporting system with different establishment numbering and some differences from State-to-State in reporting requirements and processing procedures. The worker coverage provisions of the UI system, moreover, also differ somewhat from State-to-State and likewise differ somewhat from social security coverage. In 1971 and earlier years, in particular, a number of States exempted many small employers (e.g., four or fewer employees) from UI coverage. (See Chapter V for a detailed discussion of the UI program.)

The UI estimates of employment are the lowest of the three series for most industrial categories for 1971. The lower UI estimates probably reflect primarily less comprehensive UI than social security coverage, particularly in service industries where small employers are common. The CWHS estimates are the largest for most industries, in part because the CWHS covers persons working

who didn't work during the March pay period covered by the UI and CBP data, but did work during some other part of the first quarter. The presence of government workers in " nongovernment" industries also raises CWHS estimates relative to UI and CBP estimates in some industries, particularly services. The CWHS, as tabulated in Table VII.1, however, counts each worker only once, whereas the UI and CBP data count a worker once for each job he may hold during the reference pay period.

There is no fully satisfactory way to quantify the various conceptual factors that contribute to differences among the employment series in Table VII.1. Nor can conceptual differences always be readily distinguished from differences that may arise from errors in reporting, coding, and processing the primary records entering into the three systems. While administrative record sets have been used to produce statistical series, confidentiality restrictions have limited attempts to use combinations of different administrative record sets. Matches between micro records from different systems would help considerably to quantify and resolve noncomparabilities between series. A match between micro records from the UI and SSA systems,

for example, could help identify inconsistencies in reporting unit definitions as well as inconsistent geographic and industrial coding. A match between SSA and IRS records could provide a link between place of work and place of residence, which would not only alleviate the place-of-measurement problem, but would provide a basis for construction of current data on commuting patterns.

C. Reporting and Processing Errors

Administrative agencies make great efforts to ensure the accuracy of information needed to administer their programs (net income reported to IRS or taxable wages reported to SSA, for example). Other information, important to statistical uses of the records, but only marginally applicable to administrative purposes (such as geographic and industrial information reported to SSA) are often imperfectly reported, checked, and processed.

An illustration of this problem can again be drawn from the CWHS. Not all information collected by SSA from individuals and employers is of equal importance for program administration.

Therefore, given limited resources available for ensuring accurate reporting and processing of information, it is logical to concentrate the greatest resources on the most important items. As a result, information which may be highly important for statistical applications, but of marginal importance for program administration, tends to receive low priority in competition for the resources needed to ensure timely and accurate reporting and processing of information. Information on the industry and geographic location of employer establishments, in particular, has, sufficiently little administrative importance that SSA has implemented only a voluntary establishment reporting plan for multiestablishment employers. And voluntary establishment reporting combined with limited resources for monitoring the reporting and processing of establishment records has resulted in CWHS geographic and industrial data that are frequently of questionable accuracy. Resources have not permitted a thorough evaluation. of reporting and processing accuracy, but some recent studies have suggested substantial data inaccuracies, particularly in the geographic indicators in the CWHS that have been used to develop work force migration and commuting estimates. Users of the employee-employer CWHS have noticed for some time that geographic

coding errors in the data files occasionally result in large, spurious movements of workers between geographic locations. More recently, as worker home address information has been added to the CWHS for selected years, a significant number of cases of workers with inconsistent work and residence location codes (locations beyond reasonable commuting range) has also been evident. SSA has investigated some of the more serious apparent problems and has documented a number of types of error. Resources have not been available to correct individual errors on a signifi-

cant scale, or even to estimate the relative incidence of various kinds of errors. There have been some recent studies, however, that provide some indication of the overall impact of geographic errors on selected types of data. The types of errors and overall

indicators of the extent of errors are reviewed below.

1. Reporting Problems

Because the SSA establishment reporting plan for multiunit companies is voluntary, some of the problems of incomplete and inaccurate geographic data in the CWHS result from conscious decisions of employers not to participate in the ERP. In general, however, large multiunit employers make some effort to divide employees into distinct reporting units; and when their failure to separate worker reports geographically would result in clear data distortions, SSA tries to provide special designations for the workers. The largest case of nongeographic reporting involves members of the armed forces who are placed in a special military category in the CWHS. In the case of private employer noncompliance with the ERP, SSA generally codes the workers involved into a special "Statewide" category.

While most multiunit employers do break their employees into more than one reporting unit, there is increasing evidence that many employers do not follow ERP guidelines completely in their

reports. Again, the best evidence concerning incomplete compliance with the ERP involves large government employers. A few State governments, for example, provide no reporting unit breakdowns of State workers-generally reporting there as if they were all located at the State capitol. Most State governments do divide State workers into several reporting units, but evidence suggests that in most cases the reporting units tend to be divided along agency rather than geographic lines-with geographic locations being. assigned to agency-headquarters or to some other centralized payroll accounting location. The agency reporting unit pattern appears also to hold for those few Federal civilian workers (e.g., temporary employees) subject to social security taxes. Currently, incomplete or incorrect Federal Government compliance with the ERP may not cause significant geographic distortion in the CWHS; but, with the advent of annual reporting and possible full CWHS coverage of Federal workers, the distortions could become major if ERP reporting guidelines are not adopted by Federal agencies.

Incomplete private compliance with the ERP is probably less pervasive than is the case for the Federal Government and State governments. Nonetheless, a wide variety of problems appears in

the "establishment" reports of multiunit private employers. A common practice, for example, appears to be some form of regional reporting that does not conform to county units as requested in the ERP. In addition, some companies appear to report part of their workers (such as production workers) by county and other workers (such as managerial staff) from a central location. For the most part, it would appear that these and other forms of chronic incomplete or incorrect compliance with the ERP result when employer payroll accounting procedures are not organized along lines that permit a convenient breakdown of individual employees by county of work, and employers supply instead those geographic or other organizational breakdowns that are most readily available in their payroll records.

In addition to chronic misreporting under the ERP, there is evidence of a variety of other temporary and continuing errors in geographic reporting. Employers, for example, occasionally provide establishment reports in which groups of workers have been interchanged or other. wise intermixed incorrectly among reporting units. Employers may change their reporting practices without ril- ing appropriate updated geographic information on reporting units to SSA. Multiunit or. single unit employers can supply incorrect

initial information concerning geographic location (e.g., supply a mailing address that differs from the actual county of work). Often, however, careful tracing of erroneous CWHS records is necessary to deter. mine whether information has been reported incorrectly to SSA or has been processed incorrectly at SSA.

2. Processing Problems

Some of the errors in the CWHS can be attributed to clerical errors in coding and processing employer reports to SSA. The tracing of individual errors in the CWHS suggests several ways in which coding and keypunching errors can affect the geographic information in CWHS files.

1. The reported county of work nay be incorrectly coded. Investigation of individual CWHS errors has revealed evidence of this. In some cases. workers residing in a particular county were coded as working in a county of the same name in another State. In other instances. workers were shown as working in another county of the same code in another State. There were also incidents of

transposition such as workers residing in county code 410 shown as working in county 140. There also appeared to be some confusion between city and county names. such as reporting units in Ada County, Idaho (Boise City) being coded to Boise County. Idaho.

2. The county of work may be coded correctly, but keypunched incorrectly. An example of this was discovered while investigating large commuting

flows which appeared between Now York and Alaska. Reporting units in New York were shown in Alaska because of similarities in State codes. Albany, New York, for example, is code 21000, which, if mispunched one position to the right, becomes 02100, the code for Haines Divi-

sion, Alaska.

3. The employer or reporting unit number may be mispunched. If this occurs and the mispunch is to a nonexistent EI or unit number, the worker will be unclassified by State, county, and industry. The number of unclassified workers in the preliminary first quarter files has risen dramatically in recent years, from 3 percent in 1971 to 7% in 1975. If the mispunch is to a legitimate EI or unit number, the workers will be coded to the wrong establishment and erroneous geographic and industrial information will likely result.

The timing of updates is another processing problem which can be important. The speed with which the employer file is updated with information on new employers and changes to established reporting practices has a bearing on both the number of unclassified and the number of incorrectly classified workers. If an employer notifies SSA of a change in reporting procedure at the same time they file a report on the revised basis, it is possible that the wage items from the report will be processed and matched to the, employer file before the employer file has been updated

with the new information.

3. Extent of Errors

While there has been no systematic study designed to quantify the importance of the various kinds of reporting and processing errors in geographic coding in the CWHS, several studies designed to develop migration and commuting data from CWHS files have indicated that the overall incidence of errors is substantial and may seriously impair the CWHS for use in such applications. A recent study comparing place-of-work codes with, place-of-residence codes in the CWHS, for example, was particularly indicative of the magnitude of place-of-work coding problems for large employers in the CWHS. This study was conducted as part of a larger SSA-BEA effort, sponsored by the Department of Housing and Urban Development, to prepare mid-decade commuting estimates. A 10 percent sample of workers from social security records was matched to an IRS mailing address file in order to obtain information on the workers' State and county of residence. This work was done at the request of the Secretary of HUD prior to the enactment of the Tax Reform Act of 1976, and with the concurrence of the Secretaries of

Commerce and Treasury. (See Chapter VIII for a discussion of the implications of the Tax Reform Act for interagency data linkages.)

The worker records were summarized by employer, State, county, and industry so that each summary record approximated an SSA reporting unit. Units with an estimated 60 or more covered workers suspected of having inaccurate or incomplete county-of-work coding were flagged on the basis of the following criteria:

1. 50 percent or more of the establishment's workers lived outside the county of work.
2. 10 percent or more of the establishment's commuters (county of residence and county of work differ) lived in a different BEA economic area.*

Only 3.8 percent of the reporting units were flagged. Those units flagged, however, accounted for nearly 36% of the workers with known commuting status and 60 percent of those identified as commuters. Even when the criteria was tightened to include only those units with 100 percent commuting ratios nd those with commuting ratios greater than 50 percent and more than 30 percent of the commuters from outside the economic area*, the file contained 11 percent of the workers and 29 percent of the

commuters. Units with 100 percent commuting ratios accounted for nearly 8 percent of the commuters. Many of these units had worker residences clustered an unreasonable distance away, indicating county-of-work errors. Approximately 13 percent of the commuters in the file were commuting between counties in noncontiguous States. Commuting rates for most counties were more than double the comparable rates from the 1970 census; and 1975 comparisons for selected areas covered in the Annual Housing Survey suggests the high 1975 CWHS based rates result primarily from geographic coding errors rather than increasing commuting rates, generally.

Geographic coding problems in the CWHS not only lead to erroneously large estimates of commuting, but they also bias upward estimates of work force migration based on the CWHS. Annual estimates of average interState worker migration rates derived from the 1 percent CWHS for the period 1964-74 range from a low of 6.3 percent in 1964-65 to a high of 10.1 percent in 1973-74. Data from the Current Population Survey suggest that the rates should be much lower, perhaps in the range of 3-4 percent. The estimated sharp rise in CWHS migration rates after 1970, in particular, contrast markedly with CPS data and suggest that declining SSA resources de-

voted to monitoring establishment reporting and geographic coding may be leading to a serious deterioration in the quality of CWHS migration data. In fact, without a substantial effort to edit and correct CWHS files, the

*BEA economic areas are county groupings centered on major urban areas and defined in such a way that inter-area commuting is usually minimal.

potential value of the CWHS as an inexpensive source of useful commuting and migration data is likely to remain largely unrealized.

4. Related Problems with Other Data

Both the Census Bureau and the UI employment and payroll reporting systems require mandatory establishment reporting by multiunit employers. These programs may also devote more resources to monitoring geographic information than SSA. Hence, it is likely that UI and CBP geographic data are more reliable than CWHS data. Many States, however, are reluctant to push multiunit employers for accurate county reports of employment and payroll in the UI program (although accurate State breakdowns are important for administrative purposes). Census. moreover, normally permits "estimates" of data items when accounting records do not lend themselves readily to the kinds of reports desired for statistical purposes. Hence, UI and CBP data may also be affected by the unstandardized establishment payroll accounting systems that appear to lead to incomplete and incorrect establishment reporting in the SSA reporting plan.

Errors in reports for particular establishments are difficult to monitor. but the Census Bureau has conducted some tests of the accuracy of geographic coding in its business establishment files. A recent evaluation study of the geographic coding in the 1972 Census of Retail Trade showed that the error rate at the place

(city) level was about 11 percent for establishments whose reports were based on administrative records. (it should be noted that these errors affected primarily data on the number of establishments located within the city. Since the establishments involved tended to be small, the impact of the coding errors on other data such as sales was less serious.) Many of the problems noted in this study were similar to the problems found in the CWHS commuting study (e.g., reporting from headquarters location), but relatively fewer problems seemed to result from combining information for several establishments and proportionately more problems were associated with the difficulty of using address information, supplied initially in administrative programs, in conducting censuses.

Often mailing address rather than actual location is the only geographic information available from administrative records and use of mailing addresses to compile geographic statistics can result in serious biases in the data, particularly for cities and other places in highly urbanized areas. And with increasing use of administrative mailing lists and mailed reports, the problems of obtaining reliable small area geographic data for organizations or

individuals are becoming more serious. Unfortunately, moreover, Federal administrators often have little reason to be concerned about either establishment reporting or the precise geographic location of the organizations or individuals reporting to them; and coordination between Federal administrators and State and local administrators (who do need reliable geographic information), and between administrators and statisticians has been inadequate to prevent a trend of deteriorating geographic coding in many data files at the same time that increasing use of geographic data is being made for such purposes as Federal fund allocation.

5. Errors in Other Information

Geographic reporting and coding errors are perhaps the most noticeable problem associated with using SSA and other administrative records of businesses to obtain data on employment, payroll. and other regional economic indicators. There is, however, also evidence of serious problems with other administrative records that are not central to program administration. As already noted, establishment reporting problems

in the CWHS create industrial as well as geographic coding errors. As would be expected, industry coding problems tend to increase as the desired level of industrial detail becomes finer. In a reconciliation study of establishments in the 1972 Economic Censuses and in the area sample portion of the Current Business Surveys, it was found that there were many differences in the SIC coding. For Retail Trade, the study revealed an 11 percent disagreement rate at the 2-digit SIC level and an 18 percent disagreement rate at the 4-digit SIC level. (Jeans, 1977, Table 6). Since most of the establishments in this study are nonemployers or small employers, the SIC code used in the Census would usually be derived from administrative records, while the SIC code in the area sample is derived from a business description obtained by an enumerator. These disagreement rates point to problems in the SIC coding derived from administrative records. The impact of the SIC coding errors on aggregated data such as sales would be less serious than the disagreement percentages cited above would indicate, since the establishments involved were relatively small. In the CWHS, the quality of information on individual characteristics such as sex, age, and race generally appears to be

of higher quality than business characteristics such as the location and industry of jobs. Nevertheless, there are problems with the demographic characteristics in the CWHS, particularly the race indicators. Applicants for social security numbers self identify their race as White, Negro, or Other, and evidence suggests that members of various minority groups which have been considered white for statistical purposes (such as most Mexican Americans) often have a tendency to erroneous-

ly designate themselves in the "Other" race category. Moreover, the tendency toward such erroneous designation may change over time in response to such factors as the strength of cultural or legal motivations to be identified as a member of a minority group.

D. Problems With the Timing of the Data

A problem which concerns statisticians with all data gathering activities is that of data timeliness. Generally, the more current the information is, the more useful it is to decisionmakers. An additional dimension is added to this problem when administrative records are involved-the statistician's lack of control over the timing of the data. Since the data are collected and processed by an administrative agency, processing for administrative purposes has a much higher priority and is done on a more timely basis than is processing for statistical purposes.

There are three major elements to the timeliness problem:

1. The promptness and frequency with which the data are reported to the administrative agency. In this regard administrative records are often superior to surveys and censuses. The data are reported under an ongoing program and are required by law. Reporting entities will generally provide the required information more promptly than they will respond to a periodic or one-time survey questionnaire.

2. The time required for the administrative agency to

process the information and make it available for statistical uses. This is where the above noted conflict in priority between administrative and statistical uses often causes long delays in the availability of the records for statistical purposes. The CWHS files, for example, are generally not available until 21/2 years after the end of the subject year. Many important statistical 'applications, such as the generation of data for fund distribution, require much more current data.

3. The time and difficulty of producing the desired statistics from the records. Administrative record files are often large and complex. Even when the data are made available promptly and only a sample is tabulated for statistical purposes, it can require a considerable amount of time and resources to produce the statistics. The 1% CWHS employer-employee file, for example, is approximately 1.5 million records per year.

E. Conclusion

While the CWHS illustrates the multitude of technical problems involved in using administrative records, it can also be used to illustrate ways in which administrative records can be improved for statistical uses, as well as the potential for such records to provide a powerful source of local area information for policy, planning, and research purposes.

Many of the problems described in this chapter could be resolved through improved coordination between program administrators and statistical users. In the case of the CWHS, such coordination could result in improved timing and accuracy through higher priorities and greater resources assigned. to the assembly of statistical files. Coordination between data producers and users could result in additional editing techniques to ensure the accuracy of data. Improved coordination could also increase the informational content of files available for statistical use, such as the addition of information on worker residences from W-2 forms (see Chapter V).

Improved coordination among different data collection programs could help to resolve many geographic and industrial coding

problems. For example, comparability between Census and SSA geographic and industrial codes is limited because the Census and SSA definitions of establishment differ and because coding for multiestablishment companies is carried out independently in the two systems. SSA requests that employers report on the basis of county-industry combinations--permitting, for example, a combined report for all the separate stores a retail chain operates in the same county. Moreover, SSA requests that employers assign their own four-digit reporting unit cod es to separate reporting units. For many programs, however, Census requires data for small (subcounty) geographic areas, so the SSEL has defined establish-ments in terms of operations at a single location and has assigned its own numeric codes for individual establishments. As a result, it is very difficult to check SSA establishment reports against Census materials for multiunit companies; and the effectiveness of joint Census-SSA efforts to maintain the SSA establishment reporting plan is thereby limited.

If, however, SSA requested that employers report on the same establishment basis and used the same codes as they do for the Census Annual Company Organization Survey, the SSEL could be used (provided SSA were granted access to the SSEL--see Chapter VIII) to

maintain the quality of geographic and industrial coding on the CWHS. If, in addition, the UI system used the same units and codes, the SSEL could be used to ensure uniform geographic coding among UI, CBP, and CWHS files.

Moreover, if establishment reporting were to become a mandatory requirement of the new joint IRS-SSA payroll ,reporting program, and if the W-3 form (establishment summary) were modified to request geographic and industrial activity information, it might be possible to eliminate some statistical forms presently required of employers and achieve both a reduced reporting burden for employers and improved statistical information.

A third type of coordination which can help to alleviate some of the technical problems is coordinated use of administrative

records. If, as previously suggested, micro records from different administrative record sets were merged, resulting statistical files would be far more useful than any of the individual record sets from which they were built. Such mergers could help to eliminate coverage gaps and resolve noncomparabilities between record sets. Improvements to administrative record data systems could have far reaching results. If the SSEL could be used to assign geographic codes in the CWHS, for example, it might be possible to code the records to subcounty levels. The W-3 records and associated summary statistics could then be used to produce current information for urban areas of the Nation. Since workers' sex, race, and age are available from applications for social security numbers, an expanded and improved CWHS would be capable of producing both economic and demographic information on employment, earnings, migration, and commutation for urban areas. This information would be useful to State ,and local governments, urban planners, researchers, and officials interested in targeting government programs to areas and populations most in need of assistance. The investments necessary to correct the technical problems associated with the use of administrative records would be

small relative to the costs of developing alternative sources for comparable information.

CHAPTER VIII

Legal Issues in the Statistical Use of Administrative Records

This chapter explores the complex issues with legal implications that arise when statisticians and researchers employ administrative records to carry out their purposes. The inquiry attempts to present a sense of the structure of law- built on statutes, regulations and formal policy which affect the activities of statisticians in both positive and negative ways, and which in

turn are affected and changed by those activities. there is an effort to relate the projects described in other chapters to this legal structure. And there is in addition an attempt to suggest the kinds of change in law which can improve the effectiveness of the statistical user of administrative records, while at the same time preserving and strengthening the administrative system in its ability to carry out its other functions.

With these aims, the first part introduces the concept of "functional separation," which is the cornerstone of current proposals for responsible expansion of the use of administrative records for statistics and research.

The first part examines the interests and needs of statisticians Which lead there to use information contained in administrative records. Section I points out reasons for the statistical use of administrative records. The concept of functional separation is developed in section 2 as an analytical tool for data usage. Statistical use and administrative use are defined, differentiated, and illustrated in section 3. along with terms that are relevant to legal issues. This leads to a formulation in section 4 of functional separation in general legislative terms as a way to realize the conceptual goal.

The second part of the chapter uses a characterization of the existing legal and administrative systems as a frame of reference, and suggests a set of organizing principles in which legal and statistical imperatives converge. Section 5 deals with the difficulties associated with the actual application of functional separation concepts to government agency records. Section 6 discusses the application of several confidentiality statutes to particular situations.

Finally in section 7, a brief summary of the chapter provides some suggestions for the future.

A. The Legal and Administrative System

1. Factors precipitating the shift toward greater statistical use of administrative records

Both the increased availability of administrative records, and the growing limitations on information obtainable directly from individuals on a voluntary basis, have precipitated a shift toward

greater statistical use of program and other administrative records.

Advanced data processing techniques and sophisticated methodologies have had both cause and effect implications for collection of data. As tools for statistical analysis of a broad range of public issues, they can extract, distill, and illuminate information from massive volumes of data.

At the same time, data processing capability acts as a catalyst in the development of social programs which develop complex and fine-tuned adjustments in mm of defined categories of participants, differential eligibility requirements, and other such variables. The interaction of technique and information leads to more highly refined standards of detail and quality of supporting data, and to rich program resources for decision making. The very existence of such a data base challenges the statistician to probe its availability and its adaptability for statistical uses. On the other hand, the fact that the content of administrative records is selected and shaped to the needs of, the particular, often very narrow, administrative use, creates built-in problems of definition and comparability for the statistician. This in turn generates

pressure from the statistician to influence the design of administrative data collection instruments.

The statistician's interest in using administrative records is precipitated by other factors as well, reflecting the growing resistance of respondents, both persons and firms, to cooperate with voluntary data collections While the relative strength and significance of the underlying causes of this growing reluctance are imperfectly measured and understood, some explanations seem relevant.

One is simply burden. Personal interviews by public opinion and survey research organizations, including the census and survey activities of government agencies, have proliferated in number, frequency and detail. This burden is generally imposed, moreover'

without any obvious compensating personal benefit to participants.

Another factor is public distrust of information gatherers, both governmental and private, and decline in confidence in the ability of survey organizations to preserve the confidentiality of information entrusted to diem. (1) This distrust is probably not lessened, moreover, by recent Federal requirements that respondents be told more fully the legal risks and consequences to there of providing information about themselves, including the extent of data sharing and the limitations on ability to assure con-fidentiality. (2)

Resistance of the public is reinforced when the growing volume of information collected through voluntary surveys is superimposed on the massive and regularly expanding volume of administrative collections, reaching more and larger segments of the general population, and making demands for detailed information from each. Where there is a quid pro quo, such as welfare payments or social security benefits, or a cost for not responding, such as tax duties or penalties, respondents provide administrative information in the required detail rather than forego the personal gain or suffer the cost. They may not be willing, however, to repeat or supplement the information to other collectors for other purposes.

These factors combine to raise concern about the acceptable level of response burden. counting both voluntary and involuntary collection, which can reasonably be imposed on the reporting public. Whether public resistance to burden is looked at as a decrease in the quality of data collected, or as an increase in the cost of maintaining a given level of quality of data, the perceived decline in cooperation is a development which has to be factored into agencies' data collection plans and budgets. Where administrative and statistical requirements for information compete, moreover, the program requirements generally take precedence.

As an alternative to the mounting of new surveys, the extraction of data from information collected by Federal agencies or their local counterparts in administering their social and economic programs has obvious appeal. Compiling administrative data in a microdata file can synthesize the response to a personal interview. Even where a "survey" of persons or firms is simulated by linking data about there from records scattered among several different programs or agencies, the cost may still be relatively small compared with the cost of conducting an actual survey. In

some instances, cost is a secondary factor, where personal contact would be difficult or even impossible because of inability to interview the necessary sample population, for example, deceased persons.

Another development that had consequences for the efforts of the statistician to compile and adapt administrative data was the emergence of the various privacy initiatives of the 1970's. Those initiatives grew out of the feeling of helplessness expressed by many persons about the dissemination of information about themselves, recorded in computerized records, then shared and used without their knowledge in ways that harmed or offended them. Thus the starting principle of the Privacy Act was the requirement that no disclosure be made without the consent of the person whose information was being disclosed. The practical imperatives of government were accommodated, however, in broad exceptions from the requirement of individual consent. Two other principles compensated for the erosion of consent. The first of these is the principle of notice to inform individuals what uses their information is put to. The second is the principle of accountability to the individual for the uses made without personal

consent. In combination these principles permitted normal use and exchange of information collected by government, subject to self-help methods of individual challenge to check abuse. At the same time, the development of a third principle was necessary. to accommodate the special needs for information which the statistician and researcher uses, while at the same time giving the individual full benefits of the primary principles of notice and accountability. That principle is the concept of "functional separation."

2. Concept of "Functional Separation"

"Functional separation" is a term which was chosen to conceptualize a treatment of records appropriate to the basic uses (or functions) for which those records are prepared and kept. Administrative uses and statistical uses have a polarity which needs to be recognized and built into the rules and procedures which control them. The uses of administrative records are individual in their very essence, as they are collected to do things to or for individual persons on the basis of those

individuals' rights and responsibilities. Statistical records we exactly the opposite. Individuals are examined. and their information collected and combined, as the individuals we perceived to belong in chosen study groups, and to be statistically interchangeable with others in those groups The method is to summarize. The individual is important in defining and characterizing the group, but the information about a particular individual is important not because d will be used to accomplish an individual result, but because the one individual is a proxy for many individuals. This difference in basic relationship of individual to ultimate use requires that the rules of treatment of statistical in-

formation be the obverse of the rules for treatment of administrative records. This set of concerns is the genesis of the

concept of "functional separation."

The issue of statistical use of administrative records has been scrutinized both from the confidentiality side by such agencies and commissions as HEW and the Privacy Protection Study Commission (PPSC) (3), and from the burden side by others such as the Office of Management and Budget (OMB) (4), the General Accounting Office (GAO), and the Commission on Federal Paperwork (5), The President's Statistical Reorganization Project also has more recently looked at both confidentiality and burden. From these inquiries has emerged a consensual view that the public will benefit from better access by statisticians to administrative sources of information. A caveat is added, that better access must be combined with better protection of statistical compilations of administrative data to prevent unauthorized use for non-statistical purposes.

In the quest for better statistical access combined with better data protection, increasing attention has been focused on the important concept of "functional separation" as it originated in the work of the Privacy Protection Study Commission, and was recommended for statutory treatment by both the PPSC and the

President's Statistical Reorganization Project. These projects both proposed mechanisms which took account of qualitative differences between program-administrative functions and statistical research functions, and established differential standards for managing the information needed by each.

These standards relate to access, use and disclosure of data. Functional separation means that a separate and distinct approach is necessary for the development of principles, legal rules and practices applicable to data for statistical use. While the principles and standards applicable to statistical use have to take into account the principles and standards which apply to administrative use of information, and in some respects are constrained by administrative rules, the rules for statistical data need not be similar or parallel to those for administrative use.

Applying the principle of functional separation, to make the rules appropriate for the function that the information serves, data cannot be mixed indiscriminately in statistical and administrative uses. Information designated for statistical use would not be available to administrators for their use except in anonymous or aggregate form, regardless of whether the data were

obtained directly through surveys or indirectly from administrative files. With that constraint, records compiled in administering particular programs can be used by statisticians without risk of breaching the rights and expectations of program participants about the intended uses of information they give. This aspect of functional separation has provoked considerable debate with compliance and enforcement officials, and is at the cutting edge of legislative proposals to provide legal protection to statistical files.

3. A language framework for legal issues.

The statutory background for functional separation is expressed in terms of privacy, confidentiality, disclosure, access, and other terms with special technical implications. In addition, in proposing different legal treatment of records based on different operational functions, the concept itself has added some terms, with particular meanings. This section is offered as a bridge between the legal framework which controls the flow of

information to the statistician, and the workplace within which the information is stored, used and transferred.

Generally administrative records mean records which contain information used in making decisions or determinations, or for taking actions affecting individual subjects of the records. Commonly the term refers to records about natural persons, although other entities may be treated by law as legal "persons," about whom decisions and actions are taken. Corporations, for instance, are fictitious "persons" whose actual being is created by law, and about which records are kept and decisions made. Partnerships and sole proprietorships likewise may have legal rights and duties which are separate and distinct from the legal rights and duties of the natural persons whom they represent, individually or collectively, or who conduct the business of the entity. To indicate a further level of abstraction, the estate of a deceased person - amounting merely to a bundle of residual claims and obligations - is a "taxpayer" under the Internal Revenue Code, and its records are subject to disclosure rules just as if the taxpayer were a living natural person. In other contexts, legal rules on disclosure might vary depending on whether the particular information refers to an individual in his capacity as a private

person, for instance, or as a business proprietor. The juncture of Freedom of Information Act (6) and Privacy Act disclosure rules with respect to a particular set of data may raise just such an issue.

This chapter deals with only one segment of the large volume of administrative records kept by public and private record keepers. The focus is on records kept by government agencies, mainly Federal, compiled principally in managing their social and economic programs. While agency personnel, law enforcement, regulatory. and other records are also administrative records in the broad sense, they have not been treated in scope of this discussion. Although there was no initial intention to exclude such classes of records, they demanded little attention.

It appeared in the course of examining statistical uses that agencies in which such records predominated were by and large neither providers nor users of general purpose statistical files built on an administrative record base. In the case of law enforcement records, in particular, both the administrative and the research records are subject to special legal restrictions limiting use and disclosure, and are not easily integrated into a pool of general purpose statistical data. there are some areas of study, to be sum, such as follow-up analysis of work history of ex-- offenders. that suggest the potential for careful merging of information from one data base to another. However, this potential is not likely to be realized in the form of general transfer between law enforcement and other types of data bases.

Finally, there are some arguments for excluding decennial census records from consideration as administrative records. since their purpose is almost exclusively statistics. They are, however, used for redistricting, calculating revenue sharing, providing genealogical data, and similar administrative types of use. They have been included in this study because of the special reciprocal relationship of Census with agencies using administrative records

in statistical operations. Census plays a focal role in acting as a broker between agencies in receiving, processing and merging administrative data that sometimes cannot be transferred directly between agencies. Resulting merged files can be purged of identifiers and made available to the source agencies for their statistical uses, and in many cases, to the general public for statistical use. Moreover, Census is drawing with increasing regularity on administrative files to help improve its intercessal estimate and to correct its undercount.

Statistical purposes describe purposes for which information about individual members of a defined study population is aggregated and presented without reference to individual identities. Statistical records may be kept, used. and published in microdata form-i.e., a collection of data items pertaining to one particular individual--to maximize flexibility for examining and analyzing the composition, characteristics, behavior, etc., of the group under consideration. Personal identifiers may be kept on microdata records for purposes of record validation and linkage, and the files may be transferred to statistical users with identifiers. The fact that identifiers are used in the statistical

process and are a necessary incident to the statistical file is often overlooked. Indeed, even the Privacy Act, which was meant to be a statute which would deal definitively with the issue, provided only for transfer of statistical records without individual identifiability. Of course, the individual identities of the persons making up the statistical group are not associated with the statistical files once processing is completed, nor are they material to the ultimate statistical results of the process.

Access, use, and disclosure. There are some subtle distinctions in the ideas of access, use and disclosure of records. "Access" to (or availability of) records suggests the right or the ability to see, hear, examine, or otherwise be cognizant of the information contained in the records. (in the Privacy Act sense, "access" has a further special meaning limited to the right of a natural person to examine record information about himself or herself.) "Use" generally refers to the purposes which can be served, or the operations which can be performed with the records by the person who has access. A basic distinction between statistical and non-statistical use is of principal concern. In this connection, the application of statistical methodology is not equated with statistical use. An identifiable person may be

singled out for any number of administrative actions-such as promotion, tax audit, and so on--on the basis of a statistical operation, such as ranking by specified characteristics. This would be an administrative use of statistical techniques, and not a statistical use. Quality assurance programs often involve hybrid uses of this sort, and are considered to make administrative rather than statistical use of the data.

An issue of use can be illustrated by experience of the Social Security Administration. An item designating the applicant's race is collected by SSA on its form application for a Social Security Number, exclusively for statistical use. The race item is not used in assigning the Social Security Number, nor is it used in making program determinations about the individual, which would be administrative uses. The information is used to draw samples and subsamples, and to describe racial composition of specified samples or groups of persons based on other characteristics. Inclusion of this statistical item in microdata records which are used for preparing tabulations showing the racial composition of a particular work force would be a statistical use. Such tabulations are occasionally requested by litigants in Equal Employment

Opportunity Commission (EEOC) and similar actions raising issues of discrimination in hiring or promotion. Use of the tabulations themselves as evidence in such litigation would not alter their basic statistical character. But an attempt to use those same microdata records to identify and characterize the race of particular members of that work force and to inform there that they were parties to a class action suit in which the tabulations were presented in evidence would not be a statistical use of the information. It would be an administrative use not in keeping with the statistical character of the data.

Finally, "disclosure" involves providing access or availability to another user, usually by transfer of records,

although it is evident that disclosure can take place also by word

of mouth. The significant overlapping--and to some extent circularity-of these concepts of access, use, and disclosure sometimes blurs the practical distinctions among them. Though they may seem artificial, however, the distinctions are not trivial in their relationship to the legal issues of administrative record use.

Confidentiality and privacy. These are terms which have been associated with a variety of meanings, both subjective and technical. In this chapter the terms have no arcane meanings, but make a rather simple and important distinction. Confidentiality refers to limitations which protect records from unauthorized access, use or disclosure. (For this purpose, "unauthorized" disclosure means without the consent of the person whose information is divulged, or without some other legal authority to disclose.) Privacy refers to the protected right of the individual not to be disturbed, or not to have intrusive invasions of his person or property. In this context, invasions include any type of personal contact made on the basis of record information. Finally, using the convention adopted in the Privacy and Freedom of Information Acts, the privacy concept is limited to natural

persons.

Functional separation. To summarize what is stated elsewhere, functional separation establishes two basic divisions among records, according to whether they serve administrative functions intended to have consequences for the individual subject of the record, or whether they serve statistical functions of studying groups. Functional separation principles allow information about individuals to flow from administrative sources to statistical uses, but not to return to administrators in a form associated with the identities of the individuals once the information has been incorporated in statistical records.

The concept of functional separation expresses an underlying principle of fairness in data use. That principle holds that actions and determinations about persons should be made on the basis of information which is used with their knowledge and consent. As long as statisticians and researchers do not use data in any individual way to affect the subjects of the information, their personal knowledge and consent may not be relevant. However, the collection and retention of individual information, and its use in generating new information by the researcher, require insulation from the decision process.

4. Options: Legislative approaches to functional separation

There are two principal approaches by which functional separation can lead to protected status for data committed to statistical uses. Both approaches can be found in some recent legislative proposals.

The first approach is to protect designated statistical organizational activities. The method is to name certain units as being qualified users of statistical data, to require safeguards for all statistical data within the controlled environment which they manage, and to impose limitations on access and disclosure. This is the design for the "Protected statistical center" which is described in the proposed Confidentiality of Federal Statistical Records Act. (7) The model for this approach is the Census statute (Title 13 of the United States Code) which limits examination and use of statistical records to employees of the Census Bureau. The difference which would be introduced by this proposed extension of the Census concept is that use would not be limited exclusively to employees of the organization which does the actual collection of

the data. Instead, under this proposal, data could be transferred among approved centers with relative ease. Since no data could be disclosed except among protected centers in a way which would permit such data to be associated with identifiable persons or business reporters, the agency which collected the information could even be ordered to transfer its data to other centers which demonstrated their need.

The second approach is to protect specified records or files, regardless of where they are physically located. The method is to designate particular data elements or collections of data elements as "statistical" (or research) records, and to place special conditions on the purposes for which the files can be used. In addition, this approach would restrict disclosures, both as to the form of records disclosed, and as to the type of authorized recipients. This approach is developed in the proposed Research Records Act. (8)

The Research Records Act would apply to research records as defined in the proposed statute. With respect to information about natural persons, this definition is somewhat broader than the statistical records included in it, except as records are excluded

by coverage in such statutes as the proposed Statistical Records Act, Census Act, etc. In another respect, the research proposal is narrower in scope than the proposed Statistical Records Act, since it would not apply to information about firms or other entities which are not natural persons. The proposed Research Records Act incorporates most of the recommendations of the Privacy Protection Study Commission to provide separate and distinct treatment and disclosure rules--functional separation--for the statistical and research records which it would cover.

The approach is also used in the proposed Statistical Records Act referred to above, with respect to files which would be designated by a Chief Statistician as "protected statistical centers". These latter conditions would be

somewhat less stringent than the conditions attaching to files in protected centers, and would include the use of protected files as sampling frames for disclosure of names and addresses of entities to contact in order to obtain additional information through surveys or interviews for statistical and research purposes. Both approaches have been considered in developing the "Standard Statistical Establishment List," and the legislative proposals to widen its availability. At present, the SSEL is a comprehensive national list of business entities, described by type of organization, size and activity codes, and associated with detailed financial and commercial information. The file is maintained by the Census Bureau, is used in identifiable form only by Census personnel to prepare tabulations which are made available to others in a form not permitting identification of particular firms or establishments. Some proposals for broadening access have recommended the first approach, described above, which would be to name the statistical units qualified to use the SSEL information both for preparing tabulations and for drawing samples of enterprises for surveys and questionnaires, and to exclude other statistical users. Other proposals have taken the second approach, making files available to responsible statistical users, strictly

limited to statistical and research applications. In addition, a third type of proposal has offered a "two-tier" compromise. This would create one level of establishment data to users in the general research community for approved statistical and research applications. A second level of information would be available only to Federal statistical agencies for their statistical use, and would contain data which is restricted from public availability because of its sensitive nature or because of its particular legally protected' sources. Proposals to broaden access to the SSEL are complicated by the fact that the file contains information which is Census information subject to Title 13 restrictions on disclosure as well as information which is tax return information subject to Internal Revenue Code restrictions. Because both laws restrict disclosure merely of the identity of a reporting unit, as well as disclosure of any information associated with that identity, the availability of information from the file is quite restricted.

B. Dynamics of Functional Separation

1. Dimensions and characteristics of the legal framework

Traditionally, a certain amount of statistical activity is associated directly with program operation, at least to the extent of tabulating classes and frequencies for measuring such variables as receipts, expenditures, program participation, and so on. Preparation of such statistical aggregates, in some cases, has been so closely linked to the programs whose records they reflect as to be regarded as an administrative function.

Expanding from that traditional base, statistical activities have commonly become functionally separate and independent of the operational aspects of the programs and program populations they examined. Satellite components operating within the governmental agencies which administer programs have continued as a routine matter to use the agency's administrative files as the source of statistical inquiry. The propriety of such use has seldom been questioned, at least within the Federal establishment. Most agency staff, indeed, would not ordinarily consider the availability of program records to in-house statisticians as disclosure at all,

although in a legal sense it may be. However, the laws have usually been silent about the conditions of such internal use. In the obverse situation, questions have arisen as to the proper extent of access which administrators can or should have to information produced by statisticians from those same administrative records which they sample and use statistically. Currently, for example, HHS's Office of Inspector General has broad statutory powers to demand information about individuals in compliance efforts. If exercised, such power could infringe on the policy of statistical units in HHS-contained in the Social Security Administration, the Health Cam Financing Administration, and the Public Health Services. including the National Center for Health Statistics--to release information to administrators only in aggregate or anonymous form. The agency's auditors and its Office of General Counsel may similarly claim broad access powers. and recognize few limitations on the uses which they may make of information, regardless of its statistical or nonstatistical source.

a. Disclosure within the agency, a broader view.

Authority for use of an agency's records by the agency's own employees for various agency purposes is implicitly assumed on a need-to-know basis, as observed above. Frequently there is no express authorization for such intra-agency disclosures, although the converse, restrictions on use or transfer, even within the agency, may be imposed by law. The Privacy Act, in contrast, provides explicitly for disclosures to the agency's own employer. While the principles of functional separation between statistical and other files are often carried out in administrative practice with respect to intra-agency use, they are less often subject to statutory treatment than are transfers for inter-agency use.

A somewhat different dimension of record availability may occur in a Department such as HHS, a conglomera-

tion of quasi-autonomous agencies administering a variety of separate programs which serve: partially overlapping client populations.

The legal definitions of Federal "agencies" are such that the term may mean either a Department or an operating component of that Department or both, depending on the particular statutory provisions being applied. Disclosure of identifiable individual data extracted from records compiled in administering one program to statisticians associated with another program administered by a different component within a Department has subtle but real legal implications.

Some disclosure anomalies can result from complex statutory matrices. For instance, the Tax Reform Act of 1976 (9) contains a provision allowing release of tax information to HEW (now HHS) from the Treasury Form W-2 for the sole purpose of processing the information for IRS (a component of the Treasury "agency") according to an interagency agreement. (10) The Treasury-HHS agreement provides that the Social Security Administration, an operating component of HAS, will do the processing for IRS. The Tax Reform Act contains another provision by which SSA can use tax information

to administer its own programs. (11) The interface of these provisions results in a paper transfer by HHS to SSA of data which HHS never actually obtains, and which HHS employees as such cannot use in identifiable form. SSA receives and processes the information for IRS purposes, and uses it as needed for social security purposes. But SSA must receive written approval from IRS to use the information HHS has released to it, before it can produce statistical tabulations, even though they involve no individual disclosures, when they are prepared for HHS purposes which are not related to the administration of the Social Security Act.

Furthermore, release of identifiable return information outside SSA to other HHS employees is not permitted. Indeed, even the Continuous Work History Sample (CWHS) microdata files from which identifiers have been purged is not released to researchers in HHS' Office of the Assistant Secretary for Planning and Evaluation (ASPE), despite the important research function ASPE performs for HHS, any more than they are released to the Bureau of Economic Analysis (BEA) or to any member of the general public. This is because of the residual difficulty described elsewhere of

stripping all possible association with identifiable business entities, even though no substantive information about those entities is divulged.

b. Disclosure to agency contractors.

The relationship between the program and the statistician who actually performs the work may become attenuated, and the issues then become more complex. For instance, an agency may wish to use information in its program records to study particular aspects of a client population. It may find that it lacks sufficient or suitable staff resources to commit to the necessary tasks of preparation and analysis. In such a case, the agency may enter into a contractual arrangement to have the work performed to its specifications by outside organizations. While the work product may be the same as that which would result if the agency relied on its own staff resources, the legal issues and relationships are different when the work is performed by outsiders. The agency must then deal with legal questions related to the disclosure of confidential information to others. These questions may involve a

variety of statutory considerations. Conditions are different for data controlled by the Privacy Act, for example, than for data controlled by the Census statute or the Internal Revenue Code. That is, the Census statute permits no one but Census employees to examine census returns.

The Census Bureau, as a result, does not employ contractors to perform surveys or analysis for it. On the other hand, the Privacy Act allows disclosure of covered records for a "routine use", and many agencies have determined that disclosure of information needed by contractors to perform their contractual duties would qualify as a routine use of personal information protected by the Privacy Act. In contrast, the Internal Revenue Code (as amended by the Tax Reform Act of 1976) has a provision requiring a particular type of agreement with a contractor to perform data processing functions with tax return information for purposes of tax administration. This provision applies to information about business and other tax-paying entities, as well as information about individual taxpayers. (12)

This provision enables IRS to use contractors to perform various functions involved in the administration of the tax laws including statistical activities of both IRS and the Treasury

Department's Office of Tax Analysis. The sections which make return information available to other Federal agencies-for example, the Social Security Administration, the Department of Labor, and the Census Bureau--however do not make any provision for redis-closure, nor do they provide for disclosure to contractors of those agencies. Thus those agencies cannot release return information to their contractors even in situations in which they normally employ contractors to assist there in administering their duties. The use of contractors is thus dependent on other considerations than the needs of the agency performing the work.

A number of files discussed elsewhere in this report have been unavailable for other agencies' projects because of this restriction. The CWHS file, which was used in the past by contractors of state agencies in unemploy-

ment insurance studies, cannot currently be used in those projects. Studies of subsidized housing performed for HUD by private organizations under contract, and pension studies performed for the Department of Labor by its contractors cannot have access to SSA earnings information classed as return information, even though SSA's Office of Research and Statistics has an interest in the findings, and would be willing to provide the needed information with proper safeguards. In an important pension project, SSA and the Department of Labor have been handicapped by their inability to use their contractors to process and merge return information. in this case earnings information obtained by SSA in its retirement and survivors program. Although both SSA and DOL had access to the necessary return information, the agencies' contractors could not be given access, and the scope of work to be performed by the contractors was substantially restricted, with jeopardy to the quality of the final product, because of the necessity to treat return information differently from other agency information in carrying out the steps of augmenting the files with earnings information.

Indeed, the restrictions prevent ORS from placing files containing return information in its own computer tape library which is maintained for it, with remote terminal access by an organization under contract to SSA.

Thus, in determining what information can be released to an agency's contractors, and in providing for the disposition of files upon completion of work which agencies contract for, careful consideration has to be given to the statutes which impinge on the relationship and affect the conditions and scope of work. Even when release to contractors is legally permissible, the agency will need to make adequate provision for safeguarding identifiable information contained in working files, and take appropriate steps for purging identifiers before the files are released for secondary analysis by others.

The nature of the provisions for purging of identifiers, destruction of records, and so on will be influenced by the statutory authority under which the contractual work is done. whether or not the contractor is "maintaining a system of records" as defined by the Privacy Act.

c. Disclosures among Federal agencies.

Agencies serve populations whose members are also covered in whole or in part by programs or activities administered by other agencies. In such cases two or more agencies may benefit from creating an enriched file which merges information about a sample of individuals extracted from the separate records of each agency. For instance, a group of social security beneficiaries might also be recipients of benefits administered by the Veterans Administration, and both agencies may have an interest in studying the combination of benefits.

Statistical matching techniques (13) may be used, of course, without any individual identification or disclosure. Thus, records of individuals can be selected from each agency's files on the basis of a set of specified attributes, (e.g. age, sex, race, marital status, etc.), and can be compiled without personal identifiers. The separate files without identifiers can be merged solely on the basis of similar attributes, thus synthesizing individual records without any effort to ascertain whether records of the same individuals were in fact merged.

It is more common, however, to create a merged file on the

basis of identifiers known to both agencies. When information about a sample of individuals known to the agency is used to create such a merged file, the procedure ordinarily involves disclosure in the legal sense from one agency to another of both identifiers and administrative record information. Depending on the form of the resulting file and the content of the source records, the process may involve a range of disclosure possibilities. Thus, there may be a one-way flow of identifiable data from a source agency to the agency performing the match, with a return flow of files containing merged records purged of identifiers. There may be a two-way flow of identifiable records between the participating agencies. Or there may be a one-way flow from each of the participating agencies to a third agency (for example to the Census Bureau) which would perform the operations of merging and "sanitizing" the files, and then return the resulting records only in anonymous form to both source agencies. This is the process used to perform match projects which combine SSA and IRS data.

Legal implications depend on the legal character of the source information, the cooperative agreements between and among the agencies, and the nature of the resultant files in term of the

potential for matching back against the program or statistical files of the participating agencies.

A technique used by ORS for releasing microdata files has been the restricted use agreement, as described in Chapter III. Files from which obvious identifiers have been removed, but which continue to have non-negligible risk of individual identification, may be released under user agreements to maintain their statistical anonymity, in entering into these agreements, users must stipulate that they cannot, and must agree that they will not make any effort whatsoever, to identify individuals in the file. These agreements have carried Social Security Act and Privacy Act sanctions for unauthorized disclosure. The CWHS files are not currently eligible for this kind of treatment, however, in view of IRS' restrictive position on release of microdata containing return information.

IRS has been engaged for several years in Freedom of Information Act litigation, seeking to refuse release of its Taxpayer Compliance Measurement Program (TCMP) files, which are files generated in microdata form from samples of income tax records to use for statistically analyzing tax audit formulas and audit selection criteria. Until the issues in that case are finally resolved, the future of the CWHS user agreement is indefinite.

The CWHS illustrates a number of use and disclosure issues. As noted in Chapter III and discussed elsewhere, the CWHS merges SSA files containing both benefit data compiled in its program operations and earnings data compiled in its wage reporting operations carried out in common with IRS, and defined in as tax return information controlled by the Internal Revenue Code. The CWHS does not contain occupation information, however, because that it not reported on the Form W-2 (formerly on the form 941) filed with IRS and processed by SSA. SSA access to return information does not include income tax information, in which the occupation data is contained. The CWHS consequently does not at present contain occupation. The CWHS files do contain geography and

industry coded from the Form SS-4 Application for an Employer Identification Number, which is regarded by IRS as a tax return. Because of the high degree of visibility of some employers on the basis of Standard Industrial Classification (SIC) codes associated with county code of their location, the CWHS data may be identifiable to employers, and consequently cannot currently be released to users who do not have access authorized by the Internal Revenue Code. A particularly troublesome complication has arisen with respect to BEA. BEA has had an ongoing association with ORS to perfect and use the CWHS files. In addition, BEA has provided user service by preparing tabulations on a reimbursable basis from CWHS files, including the 10 percent file which has never been publicly available in microdata form. Under the 1976 Tax Reform Act, however, BEA was given access only to corporate return information. Since the CWHS contains noncorporate employer codes for geography and industry, it cannot be provided even to BEA for analysis. This arrangement was beneficial not only to BEA and to outside users, but also to ORS, because it conserved the limited SSA resources available for servicing reimbursable requests and gave ORS the benefit of BEA's editing and improvement of the file, which invariably develops from familiarity with a file.

Another difficulty associated with the Form SS-4, Application for an Employer Identification Number (EIN) is the business birth and employer listings which were available in former years to other Federal statistical users. The Department of Agriculture can no longer make use of the SS-4 file to select a sample listing for its farm surveys.

The Department of Agriculture currently would benefit from use of the SS-4 file as a sampling frame for energy surveys, and is unable to obtain such access.

The potential value of the SSEL for statistical use is described elsewhere in this report, and cannot be overstated. Broader access, at least at the Federal level, is regarded as a necessity by most contributors to this report, and many consider that public availability to statistical users would be desirable. The legal impediments to broader access are numerous and complex. SSEL is currently compiled under Census Title 13 authority, with information supplied by SSA and IRS subject to the same disclosure restrictions as the information furnished by Census. Proposals have been under consideration since 1972 for legislative changes to broaden access. One suggestion is that name and address

information, together with industry and size codes, might be made more widely available than at present, but that other information in the file would retain Title 13 restrictions on release. One of the issues raised by this proposal is that the name. address, industry, and size code information has tax return character, at least at the time of original acquisition by Census, and its availability outside census would require changes in the Internal Revenue Code restrictions on return information. As a sampling frame, the SSEL has various advantages, but from the access standpoint. the difficulties are similar to those discussed in connection with SSA's Form SS-4 application for a Social Security Number.

It may be observed that the complexities increase if the cooperating agencies include both a Federal and a State counterpart agency, with legal consequences under both Federal and State law. For instance, difficulties are attached to the use by BLS of information provided by states from their UI reports, which contain EIN's and other information from the Form SS-4. The problems and their solution are not well defined at the present time. but it appears certain that Federal-state access conditions with respect

to return 'information will be reviewed by IRS.

One of the significant conclusions reached by the Privacy Protection Study Commission, in this connection was that States should be encouraged to insulate statistical and research records from non-statistical uses. For this, PPSC urged enactment of State statutes following PPSC policy guidelines, parallel to its recommendations for Federal records. applying the principles of "functional separation."

d. Use by non-statisticians of statistical files compiled from administrative source records.

Statistical analysis selects a small population segment to serve as proxy for a larger target population, focusing on salient characteristics, behavior, relationships, etc.

The statistical files and their analysis may, by their design or purpose, provide important information to program administrator's, oversight agencies, legislators, auditors, and courts. When these users are satisfied with statistical results based on anonymous or aggregated data, the purposes of the statistician and the non-statistician are compatible, and the statistician can conscientiously make the files available even though the ultimate uses are foreign to his own intended purposes.

Often, however, the administrator, auditor, or regulatory or enforcement officer wishes not only to use statistical results to identify population segments in which he is interested, but wishes also to locate and take action affecting individuals in the group thus identified. (The epidemiological researcher may have a similar design, though for what may be regarded as more benign purposes.) Here the objectives of the statistician are thwarted. Such uses raise doubts about the objectivity of the statistician, the premise of confidentiality on which he bases individual data collection, and the essential fairness of permitting the statistician to have free access to otherwise confidential information provided by

persons for purposes associated with their participation in particular programs.

Moreover, the statistician's sample is usually selected on attributes not associated with the action purposes of the non-statistical user, and the sample data may selectively preserve data about individuals in the sample population which are no longer retained in the underlying program files. The marriage of information from the separate files may also generate new information which was not itself contained in either of the source files. For example a level of income reported by the records in one file which is legally inconsistent with eligibility for benefits whose payment is reported by records in the other file. There are .numerous other possibilities. For example, records in a drivers' license register could be linked with records in a file containing benefit information about blind disabled persons. Of course, a "hit" (a match indicating that a particular individual has records in both files) does not automatically mean that a law has been violated. One of the listings may be erroneous; or a blind individual may have retained a drivers' license for identification to cash checks. However, discovery of such matches

may suggest abuses of some sort. Similar discoveries may attach to information about earnings which would be inconsistent with, and require disallowance of certain pensions or unemployment benefits, if the pertinent records were matched on an individual basis.

The Internal Revenue Code, for instance, contains a requirement (section 7214(a) (8)) that Treasury employees report information about taxpayer noncompliance. If this duty applies to information acquired in the course of performing statistical studies, it clearly cannot be reconciled with the functional separation principle of insulating individually identifiable statistical files from administrative actions.

Such possibilities raise ethical issues which are beyond the scope of this paper. The statistician takes the general position, however, that the administrator or enforcer ought to have access to aggregate information only, and not to individual data which has been matched for statistical purposes.

2. A closer look at some Federal statutes affecting statistical use of administrative records and protection of statistical records from nonstatistical use

In general, Federal statutes which have provided confidential treatment of record information have, by providing essentially equivalent treatment to administrative and statistical records, had a dampening effect on productive statistical efforts. For the most part, the laws have discouraged harmless interagency disclosures of identifiable data for statistical purposes at least to the same degree that they have impeded administrative disclosures, and probably more than they have impeded enforcement transfers. They have neither assisted the statistician in gaining access to program records, nor protected the record subjects from administrative actions based on statistical records.

An exception is the Census statute, Title 13 of the U.S. Code, which gives the Census Bureau broad authority to obtain information, including data contained in agencies' administrative records, at the same time it protects the Census records from being disclosed either voluntarily or by compulsion in a form which makes individual identification possible. The Census statute makes no distinction between information about natural persons and information about business entities, with the result that Census does not

ordinarily publish micro-data records about businesses which are compiled under its Title 13 authority, even though the information content itself may be publicly available from other sources. A literal reading of Title 13 prevents disclosure of any information collected under its authority and as a consequence, strict suppression procedures are applied to assure that no given item of information can be attributed directly or by inference to an identifiable respondent.

The Federal Reports Act [13a] is a record management statute which applies to solicitation of information by Federal agencies from ten or more respondents. Because of its restrictive provisions on interagency transfer, it is not an effective mechanism for authorizing transfers of identifiable data for statistical purposes. Under its provisions, statistical data can generally be transferred only in anonymous format, unless the requesting agency either has consent of each record subject, or has the power,

supported by criminal sanctions, to compel the public to provide it with the pertinent information: Such power is exceptional, particularly with reference to information for statistical use.

Some recent statutes which have been enacted to protect privacy and confidentiality of information collected by the Federal government have dealt with statistical information in ways that still frustrate legitimate statistical needs. Unless the individual consents to the disclosure, the Privacy Act of 1974 prohibits any disclosure of identifiable information except to specified classes of recipients. Statistical information can be disclosed only in a form which does not permit individual identification. Under this provision by itself, no administrative file linkage in identifiable form would be possible for statistical purposes except within the agency which collected all the information in the files to be matched. The Privacy Act basis on which agencies have disclosed data for statistical use is the provision that allows disclosure for a "routine use" which is

"compatible" with the purpose for which the agency originally collected the information. Under this provision, some administrative file linkage is performed by agencies which have joint statistical interest in the merged records, and which demonstrate compatibility of agency purposes in order to warrant the necessary disclosure. The Social Security Administration and Treasury, for instance, have created some match files which merge demographic, earnings, and income tax information for a sample of individuals whose records are contained in both agencies' administrative files. Once these files are created, they are purged of explicit identifiers, and are used in anonymous form for analysis by both agencies and by Congressional oversight committee staff. With additional suppression of information (such as geography or extremely high income level) which might lead inferentially to identification of some individuals, a public use microdata version can be produced. [14]

As noted elsewhere in this report, the Tax Reform Act of 1976 has placed stringent restrictions on the disclosure and use of information collected by IRS from and about taxpayers, both individual and business or institutional. Information about

earnings and withholding subject to the Social Security Act, including self-employment earnings, is defined by the Tax Reform Act to be within its scope. As such, it is governed by the confidentiality provisions of the Internal Revenue Code, which provides expressly but not generously for statistical applications, and which does not allow discretion for disclosure to statistical agencies not named in the statute. As described elsewhere in this report, these provisions have caused serious obstacles to many useful applications of files such as the CWHS and the SSEL.

Other statutes of record confidentiality and statutory treatment of statistical data tend to be piecemeal in coverage and rather arbitrary in scope:

Data collections sponsored by the Department of Justice Bureau of Statistics, formerly the Law Enforcement Assistance Administration (LEAA) for research, including statistical compilations, are protected from compulsory disclosure, and are permitted to be disclosed voluntarily to other researchers in accordance with LEAA approved transfer agreements. These statistical records may be compiled from administrative files such as arrest and conviction records, and obtain protection as a func-

tion of LEAA funding. [15]

Certain statistical files compiled under HHS drug treatment research authority may likewise acquire immunity from compulsory disclosure, either on the basis of their funding, or on the basis of their designation by the HHS Secretary.[16]

With opposite effect, some legislative and policy initiatives create pressures for greater statistical use, but also for administrative use of statistical findings. The National Center for Health Statistics (NCHS) for instance not only has a mandate to continue its statistical activities, but is also directed to be the central force for expanding epidemiological studies in environmental and occupational exposures to harmful substances. This mandate also includes a duty to locate and inform persons who have been exposed, and to assist there in obtaining treatment. Related to this, the National Institutes of Occupational Safety and Health (NIOSH) was provided an exception from IRS confidentiality rules in order to locate individuals found to have been exposed to known hazards. (26 U.S.C. 6103).

The Freedom of Information Act (FOIA) makes a somewhat jagged cut across these various disclosure provisions. Information about natural persons which is covered by the Privacy Act, for instance,

must be disclosed under FOIA unless its disclosure would be a "clearly unwarranted invasion of personal privacy," or unless it is protected by another statute, such as Census Title 13. FOIA requires that information about business firms must be disclosed unless its disclosure would breach trade secrets or reveal confidential financial information, or unless the disclosure is prohibited by another statute, such as the Internal Revenue Code. Other statutes interact similarly with FOIA in various patterns of inconsistency, insofar as the substantive content of the files is concerned.

In addition to statutes, government agency regulations or guidelines may complicate statistical applications based on administrative records. The Office of Management and Budget recently published guidelines under its Privacy Act authority, applicable to Federal agencies, record matching activities for purposes of fraud detection. [17] These guidelines also apply (somewhat less string-

ently) to matches for purposes other than antifraud enforcement. Although the guidelines do not prohibit file linkage, they do require reporting to Congress and OMB in advance of any matching activities. there is an exception for matching of files within an agency, for statistical purposes, but it is by no means clear whether agencies must give prior notice of planned interagency matches derived from administrative files for statistical analysis. Similarly unclear is the status of user files which are provided to agencies to identify sets of individuals for whom record information is to be extracted and matched to augment user information in a file which the agency is asked to create in order to prepare specified statistical tabulations.

More recently, OMB and EEOC have jointly published a notice of proposed guidelines for the collection of race/ ethnic data on application forms. In their present language, those guidelines permit the collection of such information, subject to their

required availability for equal rights compliance. Social Security's Application for a Social Security Number (Form SS-5), collects race/ ethnic data on a voluntary basis for statistical use, but prohibits its disclosure in identifiable form for non-statistical use, thus permitting release only in summary form for compliance purposes. While these principles of collection and disclosure need not be in conflict, considerable care and sensitivity will be needed to assure faithful treatment of confidential information provided, for statistical purposes, as well as effective pursuit of affirmative action goals.

C. Summary and Directions for the future

The discussion makes clear that the legal issues associated with expanding statistical use of administrative records are complex, often changing, and sometimes inconsistent in their results. Some insights are possible when the legal issues are examined as questions of access, use and disclosure of records. From that starting point the emerging principles can be related to

privacy and confidentiality as key concepts underlying those principles, and as embodied in legislative efforts to achieve functional separation.

The current Administration's Privacy Initiative and the President's Statistical Reorganization Project have made recommendations leading to legislative proposals for functional separation which would create quite different mechanisms for the protection of statistical records than for protection of research records, as the terms are defined in the respective proposals. Nevertheless, the line of demarcation between statistical and research records and uses is not an obvious one, and the two bills would interact to produce a complicated matrix of criteria. These legislative proposals are complex and need careful thought for the full implications to collectors and users of statistical information in specific applications. In general, however, their dual thrust is first to establish conditions permitting freer availability of information among agencies for statistical use, including agency access to Census records, and then to protect files from being used for individual actions and decisions, once the information in them has been compiled and designated by statisticians for statistical use. These broad goals are of great

importance to the work of statisticians both inside and outside the government agencies which maintain administrative files.

D. Notes and References

(1) See, for example: Louis Harris and Associates, Inc., and Dr. Alan F. Westin, "The Dimensions of Privacy: A National Opinion Research Survey of Attitudes Toward Privacy, conducted for the Sentry Insurance Co., 1979.

(2) Privacy Act of 1974; 5 U.S.C. .552a(e).

(3) Personal Privacy in an Information Society. Report of the Privacy Protection Study Commission, July 1977. Chapter 15, "The Relationship Between Citizen and Government: The Citizen as Participant in Research and Statistical Studies."

(4) Office of Management and Budget (OMB) Circular A-46.

(5) Report of the Commission on Federal Paperwork, 1978.

(6) Freedom of Information Act, 5 U.S.C. .552.

(7) The Administration's proposed bill "Confidentiality of Federal Statistical Records," is based on the recommendations of the

President's Statistical Reorganization Project, and was circulated by OMB for agency review in mid-1979.

(8) Privacy of Research Records Act of 1979, and administration bill, based on recommendations of the Privacy Protection Study Commission and the President's Privacy initiative.

(9) The Tax Reform Act of 1976, P.L. 94-455, 94th Congress. October 4. 1976.

(10) 26 U.S.C. .6103(1) (5)

(11) 26 U.S.C. .6103(l) (1)

(12) 26 U.S.C. .6103(n)

(13) Radner, D., "Report on Exact and Statistical

Matching Techniques," to be published in 1980 as an Office of Statistical Policy and Standards (OFSPS) Working Paper.

(13a) Federal Reports Act, 44 U.S.C. .3.501-3511

(14) DelBene, L., 1972 Augmented Individual Income Tax Model Exact Match File, Report No. 9, Studies from Interagency Data Linkages, 1979.

(15) 28 CFR Part 22, implementing section 524(a) of the Omnibus Crime Control and Safe Streets Act of 1968, as amended, 42 U.S.C. .3371.

(16) 42 U.S.C.. .242m, .4582; 21 U.S.C. .1175

(17) Office of Management and Budget (OMB) "Guidelines for the Conduct of Matching Pro.grams," Federal Register, March 30, 1979.

Alexander, L. and Jabine, T. "Access to Social Security Microdata Files for Research and Statistical Purposes." Social Security Bulletin. August 1978.

_____, with Scheuren, F. and Yohalem, M. "The 1976 Tax Reform Act and the Statistical Program of the Office of Research and Statistics," working paper prepared for the Subcommittee on Oversight of the House Ways and Means Committee.

Alvey, W. and Aziz, F. "Mortality Reporting in SSA Linked Data: Preliminary Results," Social Security Bulletin. November 1979.

Bateman, D. V. and Cowan, C. D. "Plans for 1980 Census Coverage Evaluation." Proceedings of the Survey Research Methods Section of the American Statistical Association. 1979.

Bounpane, P. and Jordan, C. "Plans for Coverage Improvement in the 1980 Census," Proceedings of the Social Statistics Section of the American Statistical Association, 1978, 12-20.

Buckler, W. and Smith, C. "The Continuous Work History Sample: Description and Contents," in Policy Analysis with Social Security Research Files. U.S. Social Security Administration, 1978.

Cartwright, D. "Major Geographic Limitations for CWHS Files

and Prospects for Improvement," Review of Public Data Use. March 1979.

Chandrasekaran, C. and Deming, W. "On a Method of Estimating Birth and Death Rates and the Extent of Registration," Journal of the American Statistical Association. March 1949, 101-115.

Commission on Federal Paperwork. Administrative Reform in Welfare. U.S. Government Printing Office, Washington, D.C., June 1977.

Commission on Federal Paperwork. Report of the Commission, U.S. Government Printing Office, Washington, D.C., 1978.

DelBene, L. "Augmented Individual Income Tax Model Exact Match File," Report No. 9, Studies from Interagency Data Linkages, U.S. Social Security Administration, 1979.

Dunn, E.S. "Review of Proposal for a National Data Center," Memorandum Report to the Office of Statistical Standards, U.S. Bureau of the Budget, December 1965.

Fay, R. and Herriot, R. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." Journal of the American Statistical Association, 1979, 269-277.

Fellegi, P. and Phillips, J. L. "Statistical Confidentiality:

Some Theory and Applications to Data Dissemination," Annals of Economic and Social Measurement, National Bureau of Economic Research, April 1974.

Goldsmith, J, and Hirschberg, D. "Mortality and Industrial Employment (1)," Journal of Occupational Medicine, 18 (1976). 161-164.

Hansen, M.H. "The Role and Feasibility of a National Data Bank. Based on Matched Records and Interviews," in the Report of the President's Commission on Federal Statistics. Vol. 2, 1971, 1-63.

Hausman, L. Characteristics of Selected Income-Tested Programs, U.S. Department of Health, Education, and Welfare, May 1977.

Huang, H. and Kasprzyk, D. An Examination of the Relative Benefits of Selected Sample Designs for the SIPP: ISDP Working paper #5, U.S. Department of Health, Education, and Welfare. November 1978.

Jacobson, L. "Worker Displacement in the Steel Industry," Policy Analysis with Social Security Research Files, U.S. Social Security Administration, 1978.

Jeane, Maxwell D. and Powell, John F. Memorandum on 1972 CBR Area Sample/ 1972 Economic Census Reconciliation Study," U.S. Bureau of the Census, October 31, 1977.

Kaluzny, R. Site Test Analysis: Characteristics of the Data Base, U.S. Department of Health, Education. and Welfare, May 1979.

_____, and Butler, J. The Effect of instrument Design on the Reporting of AFDC and SSI Income: A Multinomial Approach, U.S. Department of Health, Education, and Welfare, March 1980.

_____, Kilss, B. and Scheuren, F. "The 1973 CPS-IRS-SSA Exact Match Study." Social Security Bulletin, October 1978.

_____, and (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. 163-194.

Kitagawa, E. M. and Hauser, P. M. Differential Mortality in the United States: A Study in Socioeconomic

Epidemiology, Harvard University Press, Cambridge, 1973.

Klein, B. Validating AFDC Recipiency from the. Site Research Survey Using a Known Sample of Recipients, U.S. Department of Health, Education, and Welfare, forthcoming 1980.

Koteen, G. and Grayson, P. "Quality of Occupation Information on Tax Returns, " Proceedings of the Survey Research Methods Section of the American Statistical Association, 1979.

Lininger, C. (ed.). Survey of Income and Program Participation (SIPP) Conference Final Report, U.S. Department of Health, Education, and Welfare, August 1979.

Mahoney, B., Ycas, M., Kasprzyk, D., and Huang, H. Trade-offs in the Collection of Income, Wealth, and Program Statistics, U.S. Department of Health, Education, and Welfare, June 1978.

Marks, E., Seltzer, W., and Krotki, K. Population Growth Estimation: A Handbook of Vital Statistics Measurement, New York: The Population Council, 1974.

_____, Jones, C., Cullimore, S, and Foster, B. "Memorandum for the Task Force on Coverage Improvement Procedures, Subject:

Proposal for Use of Nonhousehold Sources for Coverage Improvement," U.S. Bureau of the Census, October 18, 1974.

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

National Commission on Employment and Unemployment Statistics. Counting the Labor Force, U.S. Government Printing Office, Washington, D.C., Labor Day, 1979.

Novoa, R. "Preliminary Evaluation Results Memorandum of the 1970 Census, No. 21, Subject: Listing

Census Coverage Through Drivers Licenses (E22-No. 3)," U.S. Bureau of the Census, October 21, 1971. Office of Management and. Budget, "Guidelines for the Conduct of Matching Programs," Federal Register, March 30, 1979.

The President's Statistical Reorganization Project, Federal Statistical System Project Issues and Options, Draft Report, November 1978.

Privacy Protection Study Commission. "The Relationship Between Citizen and Government the Citizen as Participant in Research and Statistical Studies," Chapter 15 in Personal Privacy in an In .formation Society. Report of the PPSC, July 1977.

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

Schiller, B., "Relative Earnings Mobility in the United States," Policy Analysis with Social Security Research Files, U.S. Social Security Administration, 1978.

Steinberg, J., Multiple Frame Sampling Approach General Framework of Alternative Approaches, U.S. Department of Health, Education, and Welfare, December 1976.

_____, Multiple Frame Sampling Approach-Proposed Design of a Pilot Test, U.S. Department of Health, Education, and Welfare, February 1977.

Stephenson, S . (ed.), Survey Research Issues Workshop: Proceedings, U.S. Department of Health, Education and Welfare, August 1978. .

Thompson, J., " 1976 Census of Camden, New Jersey Results Memorandum No. 15, Subject: Primary Results of the Camden Nonhousehold Sources Coverage Improvement Program," U.S. Bureau of the Census, October 28, 1977.

_____, "1976 Census of Travis County Results Memorandum No. 34, Subject: Travis County Nonhousehold Sources Program," U.S. Bureau of the Census, December 8, 1977.

_____, " 1976 Census of Camden,- New Jersey, Results Memorandum No. 24, Subject: Additional Results of the Camden Nonhousehold Sources Coverage Improvement Program," U.S. Bureau of the Census, October 25, 1978.

_____, "The Nonhousehold Sources Coverage Improvement Program." Proceedings of the Social Statistics Section of the American Statistical Association, 1978,435-440.

U.S. Department of Commerce. Bureau of the Census. "Infant Enumeration Study: 1950 Completeness of Enumeration of Infants Related to: Residence, Race, Birth Month, Age and Education of Mother, Occupation of Father," Procedural Studies of the 1950 Census. No.. 1, 1963.

_____. Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960. The Employer Record Check, Series ER60, No. 6, 1965.

_____. 1970 Census of Population and Housing: Evaluation and Research Program: Test of Birth Registration Completeness 1964 to

1968, PHC(E)-2, 1973a.

_____. "1970 Census of Population and Housing: Evaluation and Research Program: Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis, PHC(E)-4, 1973b.

_____. 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. 1973c.

_____. The Standard Statistical Establishment List program, Technical Paper 44, January 1979.

U.S. Department of Commerce. Bureau of Economic Analysis. Regional Work Force Characteristics and Migration Data: A Handbook on the Social Security Continuous Work History Sample and Its Application, 1976.

U.S. Department of Commerce. Office of Federal Statistical Policy and Standards. "Report on Exact and Statistical Matching Techniques," Statistical Policy Working Paper 5, 1980.

U.S. Department of Health and Human Services. Social Security Administration. Annual Statistical Supplement series to the Social Security Bulletin.

_____. LASS Working Notes Series, Nos. 1-7, 1979.

_____. Statistical Uses of Administrative Records with Emphasis on Mortality and Disability Research (Selected papers given at the 1979 Annual Meeting of the American Statistical Association in Washington, D.C.), October 1979.

Vaughan. D. Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey, U.S. Department of Health, Education, and Welfare. August 1979.

Westin, A. F. "The Dimensions of Privacy: A National opinion Research, Survey of Attitudes Toward Privacy." A Louis Harris and Associates, Inc. Survey conducted for the Sentry insurance Co., 1979.

Wittes, J. "Applications of a Multinomial Capture/Recapture Model to Epidemiological Data," Journal of the American Statistical

Association, March 1974, 93-97.

Woltman, H. and Smith, W. Preliminary Finding on Dual vs. Triple System Estimation, Internal U.S. Bureau of the Census memorandum, June 4, 1979.

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.

Ycas, M. An Introduction to the Income Survey Development program. U.S. Department of Health. Education, and Welfare, August 1979.

*U.S. GOVERNMENT PRINTING OFFICE: 1981-327~698/7103

1. Report on Statistics for Allocation of Funds GPO Stock Number 003-005-00178-6, price $2.40.

2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques GPO Stock Number 003-005-00177-8, price $2.50.

3. An Error Profile: Employment as Measured by the Current Population Survey GPO Stock Number 003-005-00182-4, price $2.75.

4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (A limited number of free copies are available from OFSPS).

5. Report on Exact and Statistical Matching Techniques. GPO Stock Number 003-005-00186-7, price $3.50.

6. Report on Statistical Uses of Administrative Records.

Copies of these working papers, as indicated, may be ordered from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.