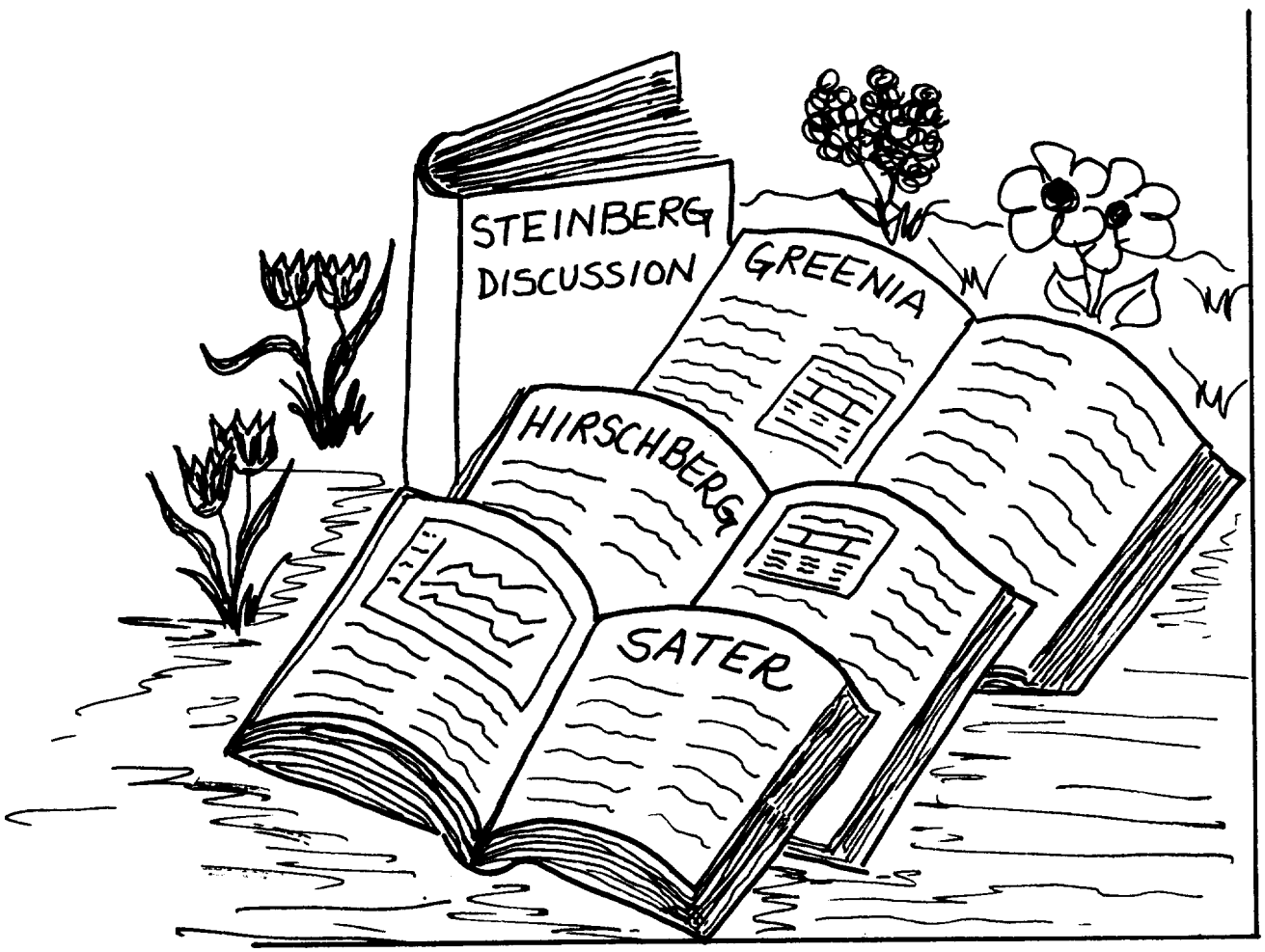
Section V:
Application
Case Studies II



1979 SOLE PROPRIETORSHIP EMPLOYMENT AND PAYROLL: PROCESSING METHODOLOGY

Nick Greenia, Internal Revenue Service

I. BACKGROUND

As the result of an interagency agreement between the Internal Revenue Service (IRS) and the Small Business Administration (SBA), IRS Statistics of Income (SOI) Division is augmenting its tabulations of business financial data (income statement, and balance sheet, when possible) with two additional data items, payroll and employment, from employment tax returns, Form 941 and Form 943. Employment is also to be used as an additional table classifier. The Small Business Administration (SBA) expects that the tabulations will prove useful in the continuing development of its Small Business Data Base in fulfillment of its Congressional mandate (P.L. 96-302 Title IV) to evaluate public policy and economic trends that affect small businesses without thereby placing any additional data collection burden on small businesses [1].

To produce these enhanced data, SOI is linking its perfected [2] sample files of business information and tax records for corporations (Form 1120 series), partnerships (Form 1065), and sole proprietorships (Schedules C, F, or Form 4835 appended to Form 1040) to their respective Forms 941 (Employer's Quarterly Federal Tax Return) and/or Forms 943 (Employer's Annual Tax Return for Agricultural Employees) in order to abstract employment and payroll from the latter two types of records. The linkage is effected through the Employer Identification Number (EIN).

These studies commence with Tax Year 1979 and will be repeated for all three types of business entity for Tax Year 1982 to coincide with the Economic Censuses. Thereafter, they will be undertaken annually for corporations and quinquennially for partnerships and sole proprietorships [3].

For the Tax Year 1979 Sole Proprietorship Employment and Payroll Study, the process entailed attempting to (a) link the 108,335 business Schedules C and F and Forms 4835 appended to Forms 1040 on the SOI Individual sample file to possible counterpart employment and payroll records in the population files of some 5 million Forms 941 and 943 for all types of business entity; (b) resolve multiple matches and mismatches for matched sole proprietorship/employment and payroll records; and (c) reweight for false unmatched sole proprietorship records.

II. SOURCE FILES

Each of the business employment and payroll studies will add employment and payroll data to the financial data already available from the IRS SOI business statistics series by matching SOI sample files of business income and tax returns with the corresponding quarterly or annual Employer's Tax Returns reporting Federal income tax withheld and Social Security (FICA) taxes (Forms 941 and Forms 943).

Processing for the 1979 Sole Proprietorship Study consisted of linking by EIN sole proprietorship business records associated with the SOI-perfected Tax Year 1979 Form 1040 sample file [4] to Census-perfected extracts of their corresponding Form 941 (Employer's Quarterly Federal Tax Return) and Form 943 (Employer's Annual Tax Return for Agricultural Employees) records. Sole proprietorship business records were appended to the sole proprietor's Form 1040 and for this study were one of the following three types:

(1) Schedule C (Profit or Loss from Business or Profession),
(2) Schedule F (Farm Income and Expenses), and
(3) Form 4835 (Farm Rental Income and Expenses and Summary of Gross Income from Farming or Fishing).

File extracts containing EIN, payroll, and employment were provided by Census for the population of some 5 million Forms 941 and 943 (Census deleted Form 943 employment due to its unreliability as a consequence of the March 12 reporting requirement, seasonality of farm employment, and exclusion of certain employee groups not under Social Security) for Calendar Years 1978, 1979, and 1980. The Census-perfected extracts of Form 941 and Form 943 data were themselves derived from tape extracts originally produced on a contractual basis by IRS (initial processor of the complete data set for tax administration purposes) as authorized by Internal Revenue Code section 6103 for Census as part of Census' ongoing effort to update annually its Standard Statistical Establishment List (SSEL).

Generally, problems of access to data were minor for SOI since all source documents were IRS-related and originally filed with IRS. While data access posed little difficulty for SOI, however, SBA could receive only tabulations of aggregated data--no files of microdata records--due to the restrictions IRS places on the disclosure of confidential taxpayer data under sections 7213 and 7431 of the Internal Revenue Code.

III. MATCH/MERGE METHODOLOGY

Foremost among the challenges presented by the 1979 Sole Proprietorship Study were those relating to the matching variable itself, the EIN, and the sole proprietorship's filing period. Each of these factors directly affected linking procedures and strategies regarding the Form 941 and Form 943 data.

While the EIN was a required entry for a Form 4835 if Form 943 was filed, it was required for a Schedule C or Schedule F if the sole proprietor had a Keogh plan (self-employed deferred compensation plan) or was required to file an employment (Form 941 or Form 943), excise, or alcohol, tobacco, and firearms tax return. Matters were complicated for Schedule C and Schedule F, however, by the Keogh plan provision

as follows. Prior to 1978, employers maintaining Keogh plans were required to have an EIN in order to complete Form 5500-K (Annual Return/Report of Employee Pension Benefit Plan for Sole Proprietorships and Partnerships with Fewer than 100 Participants and At Least One Owner-Employee), even if the only participants were owner-employees (sole proprietors and certain partners). In 1978 and 1979, owner-employee Keogh plans without common-law employee participants (i.e., with only owner-employee participants) were no longer required to file Form 5500-K, but Schedule C and Schedule F instructions for EIN completion still read as described above; that is, Keogh plans without common-law employees were not excluded explicitly. Of the more than 650,000 Forms 5500-K filed for Plan Year 1977, some 450,000 were for plans without common-law employees. Therefore, while it is unclear what the impact of such a situation was for 1979 Schedules C and F, it is apparent that the potential for problems in the 1979 Sole Proprietorship Employment and Payroll Study (false matches to Forms 941 and Forms 943) was considerable.

The EIN potential problem was compounded by the fact that while sole proprietorship Forms 941 and 943 were processed by IRS and posted by EIN to the IRS Business Master File (the computer data storage system from which the original Form 941 and Form 943 file extracts were produced for Census processing/perfection), the sole proprietorship records (Schedules C and F, Form 4835) were processed with the appropriate Forms 1040 and posted to the IRS Individual Master File (IMF) by the Form 1040's Social Security Number (SSN). Little testing or perfection was performed for the sole proprietorship's EIN, and thus, the potential for false matches as well as false non-matches--due to incorrect and even missing EIN's on the IMF side--was significant.

If the sole proprietorship's EIN posed a problem for the link operation, so did its filing or accounting period. Since (a) no such item existed on the business records themselves (it was abstracted from the one Form 1040 to which multiple sole proprietorship records could be appended), (b) a Form 1040 whose accounting period ended in other than December was presumed to have a full-year fiscal accounting period, and (c) 98.6 percent of the 92,694,302 Forms 1040 processed for Tax Year 1979 had Calendar Year 1979 accounting periods, SOI decided that part-year records and other possibly out-of-scope records (e.g., certain prior-year returns) would not be excluded from processing. Instead, the assumption was made that all sole proprietorship records should be treated as full-year calendar 1979 accounting period records. Accordingly, significant savings of both time and money were realized by disregarding the accounting period from the SOI Form 1040 sample file and using only the 1979/1980 Census Form 941/943 file for this study (instead of both the 1978/1979 and 1979/1980 files, as was done for the 1979 Partnership Employment and Payroll Study).

Since EIN generally was required as an entry on the business schedule only in the event of payroll taxes (Forms 941 and 943) or a Keogh plan, EIN-linkages could be contemplated for just a subset of the sole proprietorship sample. In fact, of the 108,335 Schedules C and F and Forms 4835 on the SOI Sole Proprietorship sample file, only 31,008 had EIN's and, therefore, could be viewed as potential initial matches with the Forms 941 and 943. By type of record, the sample counts were the following.

Form 4835:	40
Schedule F:	2,612
Schedule C:	28,356

IV. PROBLEMS AND RESOLUTIONS

Of the 31,008 records with EIN's (see Figure 1), 24,153 matched on EIN with Forms 941 and/or Forms 943 on the 1979/80 Census extract (EIN was unique for each Form 941 or Form 943 but could have been shared by a Form 941 and a Form 943). Of these 24,153 matches, 4,503 were multiple matches, meaning an SOI sole proprietorship record matched to a Form 941 or Form 943 matching either another SOI sole proprietorship record, an SOI partnership record, or an SOI corporation record. Of the inter-business entity (instead of intra-business entity) multiple matches, 117 were for sole proprietorships matching Forms 941/943 with records on either the SOI Partnership sample file or the SOI Corporation sample file. Consequences would have been dire indeed had all these multiple matches not been individually reviewed (an operation to be treated as obligatory, given the size of the largest possible sole proprietorship weight--over 2,000--and the simply astronomical amounts of payroll, hundreds of millions of dollars per Form 941 for a number of cases, reported for what were probably large corporations).

Figure 1. 1979 Sole Proprietorship Employment and Payroll
Preliminary Unweighted Processing Counts
(Pre-Reweighting)

Item	Number of Businesses (Schedule C and F, Form 4835)
Statistics of Income Sample.....	108,335
Without EIN.....	77,327
With EIN.....	31,008
Initially matched on EIN to 1979/80 Form 941 and/or Form 943.....	24,153
Initially unmatched on EIN to 1979/80 Form 941 and/or Form 943.....	6,855

All multiple matches were manually reviewed using one-line record listings containing the following data items: EIN; sole proprietorship industry code; sole proprietorship code (to distinguish between Schedules C and F and Form 4835); Form 1120/1065 code (to identify inter-

business multiple matches, but only those from SOI sample files); sole proprietorship business receipts, business deductions, and proxy payroll (salaries and wages plus cost of labor); Form 941 calendar 1979 payroll; Form 941 calendar 1980 payroll; Form 943 calendar 1979 payroll; and Form 943 calendar 1980 payroll.

At least two factors (other than the questionability of the sole proprietorship's EIN) are responsible for exacerbating the multiple match (as well as the false non-match) situation. The first is the sole proprietorship/corporation "connection" and helps explain at least some of the sole proprietorship/corporation multiple matches and mismatches. Apparently, sometimes a corporation such as a large department store will subcontract work to a sole proprietorship, say, for appliance repair or upholstery cleaning, and the sole proprietorship will incorrectly report the corporation's EIN instead of its own. The second factor concerns multiple sole proprietorships run by the same sole proprietor, even in different business activities. The sole proprietor might legitimately file several different business returns--each with the same EIN (when EIN is necessary)--and either one Form 941 or Form 943 for all businesses or one for each (also using only one EIN). Regardless, IRS would end up processing several business returns but only one consolidated (by either the proprietor or IRS) Form 941/943 containing all employment and payroll data for the sole proprietor. This latter consideration turned out to be quite significant due to the high number of "multiple matches" which were of this variety.

Resolution of multiple matches was accomplished first by "transcribing to unmatched status" sole proprietorship records with non-zero proxy payroll (the sum of salaries and wages plus cost of labor) which matched to a Form 941 or Form 943 whose payroll was egregiously greater than the sole proprietorship's proxy payroll (often sole proprietorship/corporation matches probably). Second, the assumption was made that for purposes of this processing stage, records with zero proxy payroll generally should become unmatched records. Finally, within each group of both like SSN's and EIN's (to ensure that "like" sole proprietorships also belonged to the same sole proprietor or Form 1040), the remaining matches of sole proprietorship records with non-zero proxy payroll were "perfected" by reapportioning the Form 941/943 payroll and employment data among the sole proprietorship records based on their share of the like group's total proxy payroll. When possible, this reapportionment scheme was applied according to the type of sole proprietorship record best corresponding to the Form 941 or Form 943. For example, if a Form 941 and a Form 943 matched a Schedule C and a Schedule F, the Form 941 data were accorded to the Schedule C and those of the Form 943 to the Schedule F. If a Form 941 or a Form 943 matched both a Schedule C and a Schedule F, the Form 941 or Form 943 was reapportioned among both schedules.

Comparison listings were used after resolution to ensure that all problem matches had, in fact, been remedied. Subsequent to multiple match processing, the final stage in mismatch or false match testing was performed: scrutiny and resolution of matches in which Form 941 or Form 943 payroll exceeded the business record payroll or proxy payroll by at least $1,000 (see Figure 2). Manual review of one-line listings for these records identified only 45 matches worth retaining; the remainder were dispatched to unmatched status via an algorithm which required Form 941/943 payroll to be strictly less (no tolerance) than the sole proprietorship's business deductions (business deductions was chosen in case proxy payroll had been reported or was "hidden" in deduction items other than cost of labor and slaries and wages) in order for the match to be kept. (The tolerance was dropped for this resolution process due to the large weights observed for a number of sole proprietorships and also because business deductions was sometimes zero.) Comparison listings were again used to verify that no anomalies slipped through processing [5].

Figure 2. 1979 Sole Proprietorship Employment and Payroll
Unweighted Match-Processing Counts
(Pre-Reweighting)

Category	Sole Proprietorship Records		
	Initial EIN Matches to Form 941/943	Retained as Match	Rejected as Match
TOTAL.....	24,153	22,279	1,874
Multiple business record matches.....	4,503	3,612	891
Form 941/943 payroll exceeded business deductions by $1,000*....	737	45	692
Records with zero 1979 Form 941/943 employment and payroll*	291	0	291
Other matches.....	18,622	18,622	0

* NOTE: Matched records meeting this condition but resolved as unmatched during other processing stage are excluded from this count.

The intent underlying both multiple match and mismatch processing was that only matches with almost certain probabilities of being "good" were to remain as matches. That is, the assumption was that possibly marginal matches were to be treated during these processing phases as "truly false" matches. The goal was to produce a solid reweighting base of good matches so that

reweighting for false non-matches based on their characteristics would be as accurate as possible. It was thought that any marginal cases would be more suitably accounted for later by those characteristics which allied them more closely with either true matches or true non-matches as a result of reweighting analysis.

V. REWEIGHTING

On a weighted basis, only 11.1 percent of the 12,329,982 sole proprietorships in the SOI 1979 population matched a Form 941/943 after resolution of multiple matches and mismatches. Since 82.3 percent of sole proprietorships did not have an EIN and only 7.4 percent of all unmatched records had EIN's, however, this statistic is not as discouraging as it might first appear. In fact, the match rate was 63.0 percent when only records with EIN's are considered.

Final problem adjustments consisted of reweighting for false non-matches [6], based on analytical tables of matched and unmatched frequencies classified by industry, Form 1040 adjusted gross income, business receipts, and proxy payroll (cost of labor plus salaries and wages). Unmatched frequencies were further broken down according to whether sole proprietorship records were with or without EIN, since imputation factors might differ considerably for these two sets.

Reweighting was more significant in terms of impact for the 1979 Sole Proprietorship Study than the 1979 Partnership Employment and Payroll Study [7] largely due to the sole proprietorship EIN problem (the EIN's potential absence and other complications as discussed above) and the distribution of unmatched proxy payroll. Of the $42.4 billion reported as proxy payroll by all sole proprietorship records (matched and unmatched), only $28.8 billion or 67.9 percent was accounted for by matched records. If proxy payroll is a good indicator of "true matchability" (97.7 percent of matched records also reported proxy payroll), it seemed that a significant portion of true matches remained to be "found," given that 27.6 percent of unmatched records with EIN's and 22.2 percent of unmatched records without EIN's also reported proxy payroll. Of course, to the extent that proxy payroll consists of contract labor or other "non-true" payroll components, it might not be such a good indicator for certain sole proprietorships--especially for proprietorships filing Schedules F but not required to file Form 943 for employees not under Social Security (see Data Limitations below). Imputation for "missing" data rather than reweighting for false non-matches might be more the issue then.

Reweighting was based upon a file of data defined differently in terms of matched and unmatched status from that of the 1979 Partnership Employment and Payroll Study. For the 1979 Partnership Study, a matched record was defined, primarily for reasons of simplicity and expediency (it was also the first of the business employment and payroll studies to be undertaken and, consequently, the first to encounter new obstacles and the attendant deadlines and cost restrictions in surmounting them), as any Form 1065 matching on EIN with a 1978, 1979, or 1980 Form 941 or Form 943 containing either employment or payroll for 1978, 1979, or 1980. This definition unfortunately allowed into tabulations some records with both zero employment and zero payroll for 1979, since they contained data for either 1978 or 1980. While this definition is being discontinued for future business employment and payroll studies, it also was not used for the 1979 Sole Proprietorship Study, even though a file containing two years (1979 and 1980) of Census Form 941/943 data was used for matching purposes. In fact, only records matching on EIN to a 1979 Form 941 or Form 943 containing employment or payroll data are considered matches--and these criteria must have been met even after multiple match and mismatch problem resolution. That is, records initially "matched" but later transformed to unmatched status as a result of resolution processing are not considered matched for reweighting and table purposes.

VI. DATA LIMITATIONS

Following are qualifications necessary to better understand the data in terms of conceptual limitations posed by slightly different terminologies employed across return forms as well as differences in data reporting requirements:

(a) Sole proprietorship proxy payroll was defined as the sum of salaries and wages plus cost of labor in order to be consistent with the definition of proxy payroll used for the 1979 Partnership Employment and Payroll Study. While this item was used primarily for purposes of comparison with Form 941/943 payroll during multiple match and mismatch processing, definitional differences between these two versions of payroll also warrant aggregate comparisons to ascertain what effect not only actual but also perceived differences had on the data.

Salaries and wages and cost of labor were available from Schedule C as the items wages (form instructions required the reporting of both salaries and wages) and cost of labor but from Schedule F and Form 4835 only as the item labor hired. All of these items should have excluded compensation of the proprietor, but since the Sole Proprietorship Study required gross payroll, they included amounts deducted for jobs or WIN credits.

Overstatement of proxy payroll may have occurred due to inclusion of payments for contract labor, such as certain janitorial, secretarial, or agricultural employees not reportable on Forms 941/943 but deducted on the business schedule, probably under cost of labor. On the other hand, understatement of payroll may have occurred if payroll were reported as commissions, legal and professional fees, repairs, other costs of sales and operations, or other business deductions. Additionally, for certain businesses in the Retail and Services industry groups, tip income would have been reportable on Form 941 but not claimed as a deduction on the Schedule C. Finally, a definition of payroll conforming more closely to the concept of total compensation might also contain contributions to both pension and profit-sharing plans and

employee benefit programs (such as health and prepaid dental insurance), though the proprietor's contributions to the latter were not specifically excluded by Schedule C instructions.

(b) For payroll, Form 941 appears to have required as reportable compensation virtually what was required in the counterpart Form W-2 and Form W-3 items; i.e., income which was taxable but not necessarily tax "withholdable." Form 943 required the reporting of all taxable cash wages to employees subject to FICA taxes, but excluded the value of non-cash items, such as food and lodging--potentially significant components of compensation for agricultural employees and also reportable on Schedule F as a deduction under labor hired. A further limitation was that reportable taxable wages were only required for workers under Social Security (thus, excluding many non-resident alien agricultural workers) and were not to exceed the FICA maximum, a little more than $22,000 for 1979 and for purposes of this study probably not too detrimental.

In addition to taxable wages, Form 941 required the reporting of all tips and other compensation to employees even if income or FICA taxes were not withheld and specifically excluded only annuities, supplemental unemployment compensation benefits, and gambling winnings--even if income taxes were withheld on these.

(c) While the Form 941/943 March 12 reporting date for employment was an obvious data limitation, it was exacerbated by the possibility of employment double-counting due to employees who worked two or more jobs with different employers filing different employment tax returns.

(d) While testing was conducted to identify possible mismatches in which Form 941/943 payroll was abnormally high, none was attempted (primarily due to time and other cost constraints) for possible false matches or mismatches in which it was too low. For the 1982 study, it might be possible to establish acceptable ranges for payroll/proxy payroll ratios by industry, geography, and certain size classes, but any such operation should be excessively circumspect, given "hidden" proxy payroll, as well as the problem with EIN's previously discussed. (For other recommended enhancements, see also section 10, Greenia, Nick, Match Group Case Study #00002, "1979 Sole Proprietorship Employment and Payroll.")

ACKNOWLEDGMENTS

For their thoughtful review of material in this report, the author thanks Tom Jabine (National Academy of Sciences), Carol Utter (Bureau of Labor Statistics), and Doug Sater (Bureau of the Census). Appreciation is also extended to Wendy Alvey and Beth Kilss for their help in editing the manuscript and to Dawn Nester and Rodney Turner for typing its many drafts.

NOTES AND REFERENCES

[1] For further information on the Small Business Data Base, see Kirchhoff, Bruce A. and Hirschberg, David A., "Small Business Data Base: Progress and Potential," 1981 Proceedings: American Statistical Association, Section on Survey Research Methods; Hirschberg, David A. and Phillips, Bruce, "Using Financial Statement Data to Evaluate the Status of Small Business," 1982 Proceedings: American Statistical Association, Section on Survey Research Methods; and Rose, Paul and Taylor, Linda, "Size of Employment in Statistics of Income: A New Classifier," 1982 Proceedings: American Statistical Association, Section on Survey Research Methods.

[2] File perfection essentially consisted of testing and resolving obvious math errors as well as data inconsistencies in each file record. Errors could have been made by the taxpayer or during a data processing stage.

[3] A more comprehensive treatment of small business employment and payroll, forthcoming from David A. Hirschberg and Bruce Phillips of SBA, will folow the conclusion of the Tax Year 1979 corporation and sole proprietorship studies. Final tabulations for these two studies were provided to SBA in July 1985.

[4] For a detailed account of the sampling scheme involved in selecting this sample, as well as other information--including tabulations--concerning this file, see Statistics of Income--1979/80, Sole Proprietorship Returns.

[5] For more details on the false match resolution phase, see Problems and Resolutions, Greenia, Nick, Match Group Case Study #00002, "1979 Sole Proprietorship Employment and Payroll."

[6] For a complete discussion of the reweighting process, including its assumptions, see Day, Charles, "Imputation Methodology, 1979 Forms 1040/941/943 Link Study," June 1985. This unpublished report is available upon request by writing to Director, Statistics of Income Division, D:R:S, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, DC 20224.

[7] See Greenia, Nick, Match Group Case Study #00006, "1979 Partnership Employment and Payroll."

THE DEVELOPMENT OF THE MASTER ESTABLISHMENT LIST

David Hirschberg, Small Business Administration

As part of its data base developmental effort, the Office of Advocacy, Small Business Administration (SBA), has developed a Master Establishment List (MEL) with over 8.1 million businesses. In creating the list, two commercially available lists were merged. The first, the Dun's Market Identifier file, contained over 4.6 million records; the second, the Market Data Retrieval file--a "yellow-page" listing--contained over 7 million records.

The MEL provides direct statistics on the number and geographic distribution of America's small businesses. It also facilitates communication with the small business sector and is a vital tool for conducting surveys and mailings to selected industrial sectors regarding governmental policy.

This paper describes the development of the Master Establishment List. First, some background is provided on existing small business files. Then the MEL is discussed, some of its uses are described and some on-going validation efforts are mentioned. The paper concludes by raising some of the policy implications of concern to SBA.

BACKGROUND

Although major progress has been made, the small business sector remains poorly documented in the Federal statistical system. Most existing Federal statistical data and administrative record sources are not adequate for assessing the impact on small business in a variety of policy analysis and decision-making areas. It is interesting to note that of the 124 pages of statistical tables appearing in the Economic Report of the President, 1985, only one is relevant to small business activity, "Business Formation and Business Failures, 1940-84." [1] (The source of this business formation and business failure data is Dun and Bradstreet.) Two other sources of information on business formation are the Bureau of Economic Analysis and the Internal Revenue Service. However, there are obvious problems in using their data as well. For example, the Index of Net Business Formation, published by the Bureau of Economic Analysis, is 114.8 for 1983 (with 1967 = 100). This growth level is sharply at variance with the number of business tax returns reported by IRS, as shown below. Furthermore, the number of enterprises has increased from 3.3 million in 1976 to 4.4 million in 1982.

The Small Business Administration, Office of Advocacy's Small Business Data Base was designed to provide more reliable information on the scope and contribution of the small business sector. This data base is drawn from commercially available data and places little additional paperwork burden on the business community. It permits the maintenance of confidentiality and provides policy-relevant data.

The first project, which is now complete, was the development of the United States Establishment and Enterprise Microdata (USEEM) files for 1976, 1978, 1980 and 1982. These files are based on Dun and Bradstreet's Market Identifier (DMI) files, which are collected for credit and insurance purposes. They have been edited, cleaned and reformatted, and are the basic centerpiece of the Small Business Data Base.

These four files contain information on business organizations that reported business activity in any one year. Each record which identifies an establishment has the following information: (1) Dun's number--this is a number assigned by Dun and Bradstreet that uniquely identifies each establishment and can be used to merge with prior-year files; (2) geographic location -- city, county, SMSA, state and zip code; (3) year business started; (4) number of employees; (5) annual sales volume; (6) Standard Industrial Classification (SIC) code; (7) parent and headquarter's city and state; (8) Dun's number of parent and ultimate parent; (9) subsidiary indicator; (10) status indicator -- single location, headquarters, establishment or branch; and, (11) manufacturing indicator -- indicates if manufacturing takes place at the location.

Table 1. IRS Business Tax Returns by Legal Form of Organization (in millions)

Year	Total	Proprietorships	Partnerships	Corporations
1967	8.5	6.1	.9	1.5
1976	11.3	8.1	1.1	2.1
1982	14.6	10.2	1.5	2.9

Source: Statistics of Income Division, IRS.

The USEEM files now contain data for the estimated 8 million business establishments which existed during the period 1976-82. For each year, annual files include approximately 5 million records. These records provide estimated employment and industry classification for establishments and firms, the start date (age), organizational status and geographic data for each firm.

These USEEM files have been linked into a longitudinal sample file, the United States Establishment Longitudinal Microdata File (USELM), enabling researchers to follow the same establishments over time. This is a primary and necessary requirement to address policy-relevant research issues. The 1984 files are currently being developed; they will later be merged with the USELM 1976-82 files.

The second project involves working with Dun and Bradstreet's raw financial statement file (FINSTAT). The FINSTAT file contains about 150,000 financial statements for 1975, but for the past few years the number has increased to almost 500,000 per year. To preserve the confidentiality of cooperating companies, all identifying information has been removed by Dun and Bradstreet. Although the file includes the major U.S. corporations, approximately 95 percent of the firms have fewer than 100 employees and 74 percent have fewer than 20 employees. By comparing these data with other sources, we are beginning to resolve the question of how well these data represent the small business community.

Finally, a major effort is underway to have data available on small business from the various statistical and administrative agencies of the Federal Government. Together with the Internal Revenue Service (IRS), for example, the Small Business Administration is supporting an effort to link IRS' business Statistics of Income files for partnerships, proprietorships and corporations with that agency's tax reports of employment and payrolls. This overcomes a significant shortcoming in the IRS files. As rich as they are for analytical purposes, there is no employment reported on business tax returns. Other projects include organizing the IRS Corporate Source Book [2] into machine-readable form and examining disclosure and confidentiality issues, particularly as they relate to business data from IRS and Census sources, so as to develop disclosure strategies for the release of microdata (data on individual firms).

THE MASTER ESTABLISHMENT LIST (MEL)

A universe list of firms and establishments is the core element of a statistical program. The Bureau of the Census uses the annual IRS business tax returns, combined with employer withholding/social security reports and multi-establishment company surveys, to develop their list of businesses with employees--the Standard Statistical Establishment List (SSEL). Multi-establishment companies of the Company Organization Survey enable the SSEL data to provide linkage between establishments and their parent firms. The total number of establishments in the SSEL in 1977 was approximately 4.3 million, compared with the 15.6 million business tax returns. Most of this difference is made up of firms without employees.

The Bureau of Labor Statistics (BLS) also prepares lists of establishments or, more correctly, tax units. Administrative records from each of the State unemployment insurance systems are compiled annually. Linkages between the establishments and their enterprises are not available. Other agencies have developed lists to meet their needs as well. An example is the Post Office/Survey Research Center Sample of Nonhousehold Mailers.

Unfortunately, Advocacy cannot use the Census, IRS, or BLS lists as the basis of its sampling frame. By law, the information in these sources cannot be disclosed. Therefore, Advocacy undertook to develop a Master Establishment List based on merging two publicly available private sources: (1) the Dun and Bradstreet's Market Identifier (DMI) file and (2) a "yellow-page" listing from Market Data Retrieval, Inc. (MDR) for the year 1981. The MDR file is compiled from 9 million entries, including duplicates, in the nation's telephone directory yellow pages. The MDR covers many of the establishments in the DMI file and also many small establishments and persons who do not have credit ratings.

Merging the DMI and MDR files involved a considerable effort, given the enormous size of these files and the absence of unique identifiers. [3] About 3.5 million unduplicated records in the MDR file were identified as not having a matching record in the DMI file. The resulting MEL file contains a total of 8.1 million firms and establishments for 1981. [4]

The coverage of the MEL is important. It is useful to compare with comparable tabulations of employment from the Census Bureau's County Business Patterns (CBP). Table 2 does this for the DMI components of the Master Establishment List.

The first two columns of Table 2 list the number of establishments identified in the DMI and CBP. As mentioned previously, the DMI file covers all establishments with Dun and Bradstreet credit ratings. This includes a small number of establishments with no employees, as well as an undetermined number of small establishments with employees. In contrast, the CBP includes only establishments with employees. Given these coverage differences, it is noteworthy that there is a basic similarity in the total number of establishments.

Several reasons exist for the differences by industry, but they are difficult to quantify. Discrepancies may result from differences in industrial classification between the DMI and the CBP. The extent to which the DMI file includes firms with no employees, as well as establishments which are no longer in business, is not known.

Given these classification and coverage problems, the employment estimates are remarkably similar at the major industry division level, as shown in Table 3. Total employment in the DMI file is 6 percent less than that of BLS and 2 percent more than that of CBP. For mining, contract construction, manufacturing, and services, the DMI reports slightly more employment

Table 2. Establishment Counts by Major Industry Division: Dun's Market Identifier (DMI) and County Business Patterns (CBP), 1981

(Establishments in Thousands)

Industry	DMI	CBP	Ratio DMI/CBP
All Industries, Total	4,635	4,587	1.01
Agriculture, Forestry & Fishery	120	804	.15
Mining	42	359	.12
Construction	612	626	.98
Manufacturing	441	336	1.31
Transportation, Communications & Public Utilities	182	162	1.12
Wholesale Trade & Retail Trade	1,846	1,887	.98
Finance, Insurance & Real Estate	372	387	.96
Services	1,019	1,445	.71

Note: Components may not add to total due to rounding.
Source: Tabulations from the DMI and County Business Patterns, U.S. Bureau of the Census (selected years).

Table 3. Employment by Major Industry Division: Dun's Market Identifier (DMI), Bureau of Labor Statistics (BLS) and County Business Patterns (CBP), 1981

(Employment in Millions)

Industry	DMI	BLS	CBP	Ratio		
				CBP/DMI	BLS/DMI	BLS/CBP
All Industries, Total	74.7	75.1	74.8	1.001	1.005	1.004
Agriculture, Forestry & Fishery	.8	NA	.3	.38	NA	NA
Mining	1.3	1.1	1.1	.85	.85	1.00
Construction	5.9	4.2	4.3	.73	.71	.98
Manufacturing	21.2	20.2	20.4	.96	.95	.99
Transportation, Communications, & Public Utilities	4.1	5.2	4.6	1.12	1.27	1.13
Wholesale Trade & Retail Trade	16.7	21.6	20.3	1.22	1.29	1.06
Finance, Insurance & Real Estate	4.6	5.2	5.4	1.17	1.15	.98
Services	19.0	18.6	17.9	.94	.98	1.04

Note: Components may not add to total due to rounding.
Source: Preliminary Report on the Development of the Master Establishment List, 1982, Social and Scientific Systems, Inc.

Table 4. Dun's Market Identifier (DMI) and Market Data Retrieval (MDR) Files as Components of the Master Establishment List, 1981

Number of Establishments in Thousands

Industry	DMI	MDR	MEL	Ratio MDR/DMI
All Industries, Total	4,635	3,488	8,123	.75
Agriculture, Forestry & Fishery	120	49	169	.40
Mining	42	10	52	.25
Construction	612	215	828	.35
Manufacturing	442	82	524	.19
Transportation, Communications & Public Utilities	182	84	267	.46
Wholesale Trade & Retail Trade	1,846	1,054	2,900	.57
Finance, Insurance & Real Estate	372	407	779	1.09
Services	1,019	1,577	2,595	1.54

Note: Components may not add to total due to rounding.
Source: Preliminary Report on the Development of the Master Establishment List, 1982, Social and Scientific Systems, Inc.

than the CBP or BLS files. However, there is significant undercoverage for wholesale and retail trade; transportation, communications and public utilities; and finance, insurance and real estate.

Unfortunately, employment is not available from the MDR file, but the number of establishments added to the DMI file is shown in Table 4. It was apparent from the detailed industry tabulations that the added MDR firms were mostly professionals, such as doctors and lawyers, as well as taxi operators, truckers, insurance agents, and real estate brokers -- businesses that generally do not use credit. These sectors are basic to small business activity and it is important that they be included in lists of small businesses.

In contrast to the 15 million tax returns filed with IRS, the Master Establishment List contains 8.1 million firms and establishments. It does not follow that there is a deficiency in the MEL. Inspection of the sales distribution reported in IRS' proprietorship files suggests that they include persons with other occupations and do not truly reflect full-time business activity. Of the 12.7 million proprietorship reports in 1980, almost half have business receipts below $5,000.

The analysis of the DMI file and the business units added by the MDR file indicate that, for most purposes for which the file will be used, the MEL is representative of the full-time business population with employees.

USES OF THE MEL

The Master Establishment List has been used for a variety of purposes. Users studying specific problems relating to small business have requested that the Small Business Administration make specialized tabulations from the MEL, draw samples based on those tabulations, and provide mailing lists for the sample cases. In some cases the requests have asked for firms by industry and size for a specific State or designated SMSAs or even particular counties. Although some users have been concerned with the broad spectrum of business units, other users' interests have been highly specialized.

An example of the use of the MEL to create a specialized data base was its use in analyzing the proposed legislation on enterprise zones. Because the establishments in the MEL have addresses, it is possible to examine the existing location of business activity in central cities and non-central cities in relation to the proposed enterprise zones. Some measure of the magnitude of potential costs and benefits of the legislation can be obtained by analyzing projected changes in business activity and employment.

In another application, using a three percent sample of the MEL's businesses, an Ownership Characteristics Survey was initiated in January of 1984. It asked respondents for the legal form of ownership as well as for the sex, race and veterans status of the business owner.

Summary results are available in the "Report of the President on the State of Small Business, 1985." [5]

VALIDATION EFFORTS

The exact matching of the 4.6 million DMI records and the 9 million MDR records to produce 8.1 million Master Establishment List records was considerably more successful than might have been expected, and the resulting MEL file has had wide use. As the tabulations of MEL show, the DMI data were augmented in precisely those areas where it was known that coverage was incomplete (i.e., services and trade). Although there are undoubtedly additional small businesses that are without Dun's credit ratings and are not listed in the yellow pages, it is not clear that further efforts to extend the MEL would be worthwhile.

Validation studies have been carried out analyzing the MEL's coverage, consistency, and completeness. One such study involved matching the establishments in the area samples of the University of Michigan's Survey Research Center with the establishments listed in source areas in the Master Establishment List. Another study is comparing State unemployment insurance data with DMI files.

The former study revealed important differences in the MEL list and the list compiled by Michigan. However, recent research has indicated that these lists are subject to obsolescence. Turnover is about one percent a month; therefore, if lists compiled for different time periods are compared, a large number of nonmatches should be expected. This and other experience has shown that a large proportion of nonmatches occurs when business lists are matched using different sources of information. [6]

In the latter study, unemployment insurance microdata files and DMI files were matched for a recent time period for Texas and Pennsylvania. When the comparisons are completed, they will yield information of considerable value in evaluating the DMI file. It can be noted that only about 40 percent of the firms in the files were matched.

FEDERAL POLICY IMPLICATIONS

Using the January 1985 DMI and MDR files, an updated MEL is being created. We are asking support from the various statistical agencies to provide resources to continue this effort, to improve its quality and help make it generally available to the statistical community.

There is a clear need throughout the Federal establishment for a consistent and reliable business universe frame for a variety of research and sampling purposes. Each Federal agency now operates its own system, virtually oblivious to the activities and requirements of others. Employment differences between systems are explained as due to classification, reporting and coverage procedures. In this time of considerable budgetary restraint, cooperation in the development of databases such as the MEL is absolutely necessary.

ACKNOWLEDGMENTS

The development of FINDIT and the actual computer matching were carried out by Social and Scientific Systems, Inc., a Maryland based 8(a) firm. The contributions of Dennis Abels, Wendy Alvey, Sam George, and Pal Khera are gratefully acknowledged.

NOTES AND REFERENCES

[1] Council of Economic Advisors. (1985) Economic Report of the President, 1985, table B-91, p. 337.

[2] The Internal Revenue Service's Statistics of Income Division produces a Corporate Source Book annually, which presents detailed income and balance sheet data classified by industry and size of total assets. For more information, contact the Corporation Returns Analysis Section, Statistics of Income Division (D:R:S:C) at IRS.

[3] For a detailed description of the methodology and computer algorithm, see "File Matching Utilizing Automated Heuristic Techniques (FINDIT)," by Social and Scientific Systems, Inc., Bethesda, MD, January 1983.

[4] See "Preliminary Reports on the Development of the Master Establishment List," by Social and Scientific Systems, Inc., Bethesda, MD.

[5] Small Business Administration. (1985) The State of Small Business: A Report of the President, 1985.

[6] Converse, Muriel and Heeringa, Steven G. (1984) "An Evaluation of the Accuracy and Current Utility of the 1981 Master Establishment List (MEL)," Institute for Social Research, University of Michigan, Ann Arbor, MI.

ENHANCING DATA FROM THE SUPVEY OF INCOME AND PROGPAM PARTICIPATION WITH DATA FROM ECONOMIC CENSUSES AND SURVEYS--A BRIEF DISCUSSION OF MATCHING METHODOLOGY

Douglas K. Sater, Bureau of the Census

This discussion involves the enhancement of data from the Survey of Income and Program Participation (SIPP) with data from economic censuses and surveys. This is a pilot project and is still in the development stages.

This discussion focuses on the matching methodology, problems, and problem resolution.

I. INTRODUCTION

The Survey of Income and Program and Participation is a new Census Bureau Survey designed to collect a host of information on the social, demographic, and economic situation of the nation's individuals and families.

The data will be extremely valuable to labor market analysis, but they have one major shortcoming--they do not include characteristics of the employer for which the sample persons worked. This gap can be bridged by the addition of information on employers that is collected in the economic censuses.

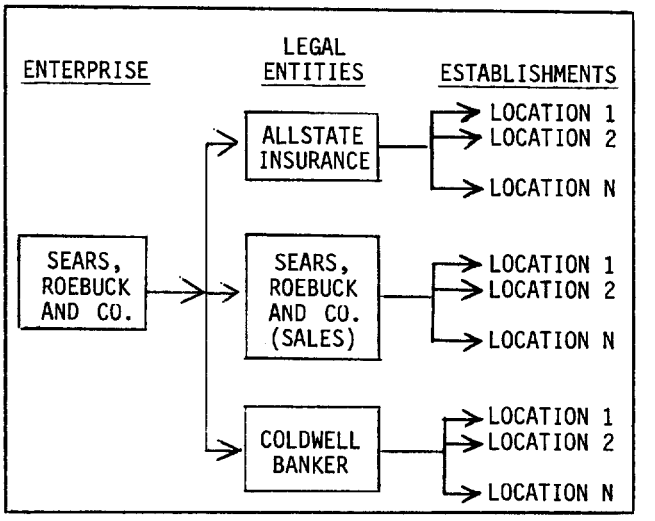
The addition of economic data to the SIPP will enable researchers to obtain improved estimates of the impact of economic and institutional forces which have been intensively studied but are only partially understood or measured. Some of the areas in which the matched file can yield new insights are: the relationship between capital and wage rates, structural unemployment, the transition from a goods to a service economy, unions and the labor market, productivity analysis and numerous other studies. For some of the studies, data at the establishment level are appropriate, and for others, enterprise level data are needed.

II. DEFINITIONS

An establishment is defined as a single physical location where business is conducted or where services or industrial operations are performed. Where separate activities are performed at a single physical location, each activity is treated as a separate establishment. The legal entity is an organizational unit which is assigned an employer identification number (EIN) by the IRS for tax reporting purposes. The legal entity represented by the EIN may comprise one or more establishments. The enterprise is the entire economic unit consisting of one or more establishments or legal entities under common ownership or control. The following figure (Figure 1) shows a partial example of these definitions.

We will be conducting the matching activity for about 20,000 persons in Wave 6 of the SIPP -- the first annual "round-up." In addition to the demographic and economic information, the Wave 6 questionnaire also asks for the employer name, address, and employer identification number for up to three employers.

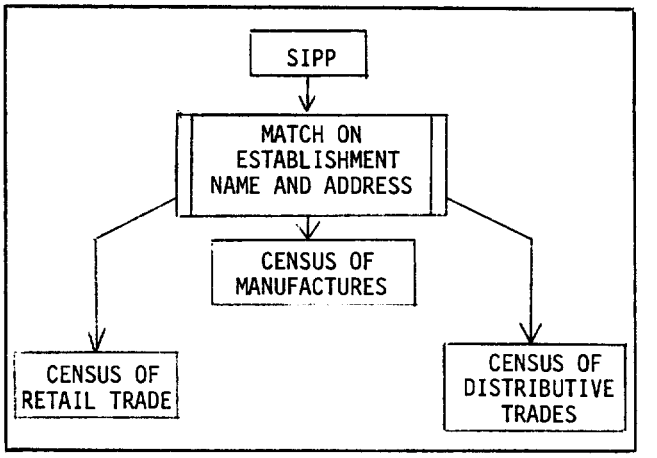
Figure 1.--A Partial Example of Basic Definitions



The first step in this process was to examine the available economic data sources. The Census Bureau conducts numerous economic censuses and surveys, such as the Census of Manufactures, which contain the needed economic data. For linkage purposes, the economic census records also contain a census file number (CFN) which uniquely identifies the establishment. They also contain the establishment name and the establishment address, but they do not contain the EIN.

The first option would be to match the SIPP directly to each economic census needed. (Figure 2 shows a simplified diagram with

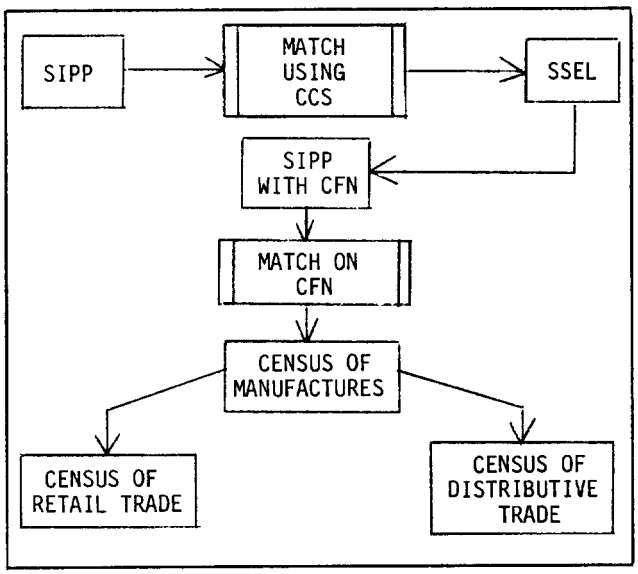
Figure 2.--Simplified Diagram of Direct Match to Three Economic Censuses



only 3 possible economic data sources.) This would involve numerous matches on employer names and addresses. Since we are only trying to match about 20,000 cases, the development and testing of programs and the sorting of the economic files were more than we wanted to tackle in this pilot project. Further, the economic censuses do not cover all establishments. That is, they do not cover some "out-of-scope" establishments nor do they cover small establishments. Since about half of all establishments have less than 5 employees, this is a serious shortfall for our purposes.

A more attractive approach would be to conduct the match through an intermediate data set and program system, namely the Standard Statistical Establishment List (SSEL) and the Census Control System or CCS (Figure 3). The SSEL is a centralized multipurpose computerized name and address file of all known

Figure 3.--Simplified Diagram of Match to Three Economic Censuses Using the SSEL and the CCS



employer firms and nonemployer agricultural firms. (This includes the out-of-scope and small establishments as well as establishments covered by the economic census.) The CCS is an interactive random access name search program and series of files derived from the SSEL. It contains the establishment name and address, the EIN and the census file number. The file also contains selected search keys: ZIP Code from the address, a name search key and the EIN. Further, these files also contain selected data such as the number of employees and the annual payroll. In essence, the CCS is a computer assisted manual search program, and it seems to fit our needs quite nicely. Thus, the approach taken is to use the CCS to match to the SSEL to pick up the CFN and selected bits of data. The CFN will then be used to match to the economic censuses. The CFN has another nice property, it allows us to match at the establishment or the enterprise level.

The CCS operates in two basic modes:

1. In the EIN mode, one provides the system with the EIN and it returns an abbreviated SSEL record for that EIN.
2. In the name search mode, one provides the system with the name. The system compresses the name, selects the search key, locates the block of records corresponding to this name key, and returns all records in this block. Additional screening is performed based on other data (such as ZIP Code) if it's provided to the system. The selection of the correct record is then done manually.

   For multi-establishment enterprises, located in either the EIN or the name search mode, a second search is done which lists all establishments within the legal entity or enterprise, as appropriate. The selection of the correct establishment record is then done manually.

A hypothetical example would be as follows: Suppose one wanted to locate American Art Supplies, 1235 Main Street, 20735. We would provide the system with "American Art Supplies, 20735".

It would return, for example, the following three records from the Block:

1. American Art Supplies
2. American Fabricaters
3. American Farm Products

We then select record (1) and it provides a second listing containing, for example, the following two records:

1. American Art Supplies-Hqt.
   1235 Main Street.
2. American Art Supplies-Sales
   425 Canal Street.

We then extract the CFN associated with record 1. This is an oversimplification of the system but it gives a general idea of the process.

To make the process as efficient as possible, a stage-by-stage process has been designed which maximizes the amount of computer work and minimizes the amount of manual review. For example, well-considered sorting of the SIPP file can greatly speed the process. That is, assembling the same employer names into groups will allow one search for many records with the same name. Employers of 250 or more employees account for less than 1 percent of all employers, but account for 31 percent of all employees.

III. MATCHING PROBLEMS

There are numerous problems with name matching. First, there are reported name variations due to abbreviations, misspellings, etc. For a household interview survey, such as the SIPP, there are several things

that must occur to get a correct name spelling. The interviewer must hear the response and spell the name when filling in the form. The data keyer must be able to read the written entry and key the name. This, in itself is more than ample opportunity for the introduction of errors. Plus, there are errors introduced through phonetic problems. Names such as KROEHLER, BEALLS FLORIST, BURROUGHS, and PFEIFFER BREWERY would pose such problems.

Also, the SSEL, as good as it is, does contain some typographic errors. At any rate, most of these cases are expected to be resolved through the computer assisted manual search process using the reported address and "judgement." For example, if we are trying to locate "KRAYLER, 75 Ely Street, Binghamton, N.Y." we might decide that this is really "Kroehler Manufacturing Co. of Binghamton." We are referring to this process of decision as "judgement" because some degree of uncertainty may exist. If the level of uncertainty seems excessive, the case will be referred for further review. However, care must be exercised in the implementation of "judgement." It implies a lack of uniformity and nonempirical matching criterion.

Another problem is the reported name variations for franchises and "Doing Business As" vs. legal name. As an example, an establishment may be commonly known as "Wendy's," but in actuality, it is a franchise using the Wendy's name and whose legal name is John Smith Enterprises. The match process does not have, in its design, an a priori process to resolve these problems, but the professional review process may be able to identify and resolve such cases.

A potential problem is the presence of mailing address on the SSEL rather than the physical address. Although every effort is made to obtain the physical address for the SSEL file, there are occurrences where the address on the SSEL is the address of the lawyer, accountant, or the administrative office. Depending on the particular circumstances, the problems may be solved or may be intractable.

Also, multiple establishment names on SSEL records may cause problems.
These are occurrences of different establishments having the same name. A hypothetical example would be as follows:

Clinton Aluminum (Hdqts.)
1235 Main Street
Clinton Aluminum (Mfg)
751 Ash Street
Clinton Aluminum (Sales)
755 Ash Street

This, in itself, poses no major problems, unless the address is not reported in the SIPP. Thus, the first question is whether there is sufficient name detail reported in the SIPP to match such a case without address? That is, are division or group names reported in the SIPP? Given the amount of space on the form, I think not. A typical SIPP entry for this example would simply be "Clinton Aluminum." In this event, other matching criteria need to be implemented. If each establishment is in a different part of the country, the selection of the establishment within the same SMSA as the SIPP respondent's may be a reasonable criterion. Another possibility would be to use the SIPP respondent's occupation. For example, if the occupation were salesman, a reasonable criterion would be to assign the case to Clinton Aluminum - Sales Division.

Suppose, in the Clinton Aluminum example, we have located the correct legal entity, but cannot match to the correct establishment. This case should not be hastily written off as a nonmatch. We already know alot about it. We know the enterprise, the legal entity, and we know that it is one of three establishments. It seems that a conditional allocation process will maximize the amount of information. There are several ideas for performing this allocation. One approach would be to use an average value for all three establishments. Another would be to randomly assign the case to one of the three establishments or to do the assignment according to a probability function based on employment size. The probability of correct match is that dependent on the probability function and, for mismatches, data utility is dependent on the degree of homogeneity of the three establishments. In the Clinton Aluminum example, suppose that all three establishments are the same size. Then the chance of a correct match is one in three. In this same example, the wage structure and degree of unionization, etc. are likely to be quite different between the establishments. Thus, a mismatch will distort the data. In a case such as Wendy's or McDonald's, such data distortion would be minimal.

I have not considered this allocation process in depth, but will in the next few months. At any rate, I will need to assign two sets of flags to keep track of what was done and how well the record was matched. The first will identify the type of match. The second will apply to allocated matches and will provide an assessment of the probability of correct match.

IV. PRE-TEST RESULTS

A small-scale familiarization test of this computer-assisted manual search process using the Census Control System was conducted. The sample was comprised of 166 employer names reported in the Waves 1 and 2 of the 1984 SIPP. These cases were drawn from a sample of Primary Sampling Units (PSU). These PSU's were not scientifically sampled, but were arbitrarily chosen to include (1) a variety of PSU's (by size and region), and (2) a variety of manufacturers. Because this is not a scientific sample and only manufacturers are included, the results cannot be generalized and are included only as an approximate indicator. The purpose of this exercise was primarily educational; that is, to see how the process works with real data.

Waves 1 and 2 asked for the name of the employer for which the person worked during the reference period. Although the employer address and Employer Identification Number were not collected in these waves, we tried to obtain the employer addresses for these cases from a variety of reference materials, such as the Major Employer Lists from the 1980 census, telephone directories, and Standard and Poor's Index of Corporations. Table 1 shows the different levels of employer information and the proportion of cases at each of these levels. Table 2 shows selected results of this test.

Table 1.--Results of Address Search Operation

Item	No.	PCT
Total	166	100.0
With Corp. Hdqts	94	56.6
No Corp. Hdqts	72	43.4
With Estab. Address	72	43.4
With Corp. Hdqts	44	26.9
No Corp. Hdqts	28	16.9
No Estab. Address	94	56.6
With Corp. Hdqts	50	30.1
No Corp. Hdqts	44	26.5

Even though an establishment address was found for only 43 percent of the cases, the employer name in the SIPP was matched to the correct enterprise 78 percent of the time. The similar match rate is 78 percent for legal entities and 51 percent for establishments. For those cases where there was an establishment address, the match rates are: 88 percent for enterprises, 88 percent for legal entities, and 81 percent for establishments. (Note that the lines "Matched to Enterprise" and "Matched to Legal Entity" are not equivalent. As an example, if a person reported he/she worked for Sears, Roebuck and Company, the person can be matched to the enterprise, but not to the legal entity. That is, which of the following would be the correct legal entity: Allstate Insurance, Coldwell Banker & Co., Dean Witter Financial Services, or Sears Merchandise group? As it turns out in this very small-scale test, we did not encounter any cases of this type. Hence, the number matched to legal entity is 130 and the number matched to enterprise is 130.)

1. Type 1 -- These nonmatches represent cases where there were more than one establishment with the same name all at different addresses. If the address was reported in the SIPP, we would have been able to match these cases. Thirty-one of the 46 nonmatch cases were Type 1's.

Table 2.--Results of Matching Test

SIPP-SSEL Match Status	Total		With Establishment Address		No Establishment Address	
	Number	Percent	Number	Percent	Number	Percent
Total	166	100.0	72	100.0	94	100.0
Matched to Enterprise	130	78.3	63	87.5	67	71.3
Matched to Legal Entity (EIN)	130	78.3	63	87.5	67	71.3
Matched to Establishment	84	50.6	58	80.6	26	27.7
Uniquely Identified by Name	75	45.2	49	68.1	26	27.7
Uniquely Identified by Name & Address	9	5.4	9	12.5	X	X
Not Matched to Establishment	46	27.7	5	6.9	41	43.6
Type 1	31	18.7	X	X	31	33.0
Type 2	9	5.4	5	6.9	4	4.3
Type 3	6	3.6	0	.0	6	.0
Type 4	0	0	0	.0	0	.0
Not Matched to Legal Entity (EIN)	36	21.7	9	12.5	27	28.7
Not Matched to Enterprise	36	21.7	9	12.5	27	28.7

X -- Data cell does not apply.
Type 1 -- These nonmatches represent cases where more than one establishment was found in the SSEL, all at different addresses (but part of the same company) and the company name matched the name reported in the SIPP.
Type 2 -- These nonmatch cases represent more than one establishment at the same address in the SSEL; that is, we would need more information than just the address (such as plant or division name or SIPP occupation) to identify the correct establishment.
Type 3 -- These are cases where the SSEL contains mixed types of entries, some Type 1 and some Type 2.
Type 4 -- These are cases where we could not identify any establishments in the enterprise by name. There were no Type 4's in the test.
(See text for more details on the definitions of the nonmatch types 1-4.)

2. Type 2--These are cases where there are more than one establishment with the same name and at the same address that is, we need more information than just the name and address (such as plant or division name or SIPP occupation). Nine of the 46 nonmatch cases were of this type.

3. Type 3--These are cases where the SSEL contains mixed types of entries, some Type 1 and some Type 2.

4. Type 4--These are cases where we could not identify any establishments within the enterprise by name. There were no Type 4's in the test.

There were 36 cases for which we could not locate the enterprise on the first pass. A large part of this is due to the lack of address for these cases. For the 16 of these, the location was apparently outside the search area we tried (PSU of SIPP respondents address). An address reported in the SIPP will permit us to match most of these. Also, we were able to locate an additional 12 through further research. These were, in general, very small companies. The remaining 8 are, as yet, unresolved. Given the nature of this test, these results were most encouraging.

The 130 SIPP-SSEL matched cass were also matched to the Census of Manufacturers (CM). Of these, 100 matched exactly 26 matched to the enterprise, but the establishment was non-manufacturing and not in the CM, 3 very small and out-of-scope for the CM, and the remaining case was a true nonmatch.

V. OTHER ISSUES

There are a number of other issues to be faced in this project, some of which are:

1. Adjustment for nonmatches--allocation or reweighting. Nonmatch rates will be significantly different between large and small employers. Since much of the analysis will be affected by this, some sort of allocation or reweighting will be necessary.

2. Development of match status flags and probability of correct match status.

3. Development of a process of computing match error rates.

4. Errors in EIN's.

5. Differences in reference periods between the Economic Censuses, SSEL, and the SIPP.

6. Suppression issues in data releases.

We will be investigating these issues in the next few months as work on this pilot project progresses.

BIBLIOGRAPHY

1. Sater, Douglas K., "Enhancing Data from the Survey of Income and Program Participation with Data from Economic Censuses and Surveys," Unpublished paper, July, 1985.

2. Haber, Sheldon E., et al., "Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study," American Statistical Association Proceedings, Social Statistics Section, 1984.

3. U.S. Department of Commerce, Bureau of the Census, The Standard Statistical Establishment List Program, Technical Paper 44, January, 1979.

4. Kasprzyk, Daniel, and Roger A. Herriot, "The Survey of Income and Program Participation," American Statistical Association Proceedings, Social Statistics Section, 1984.

5. U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Statistical Policy Working Paper 2-- Report on Statistical Disclosure and Disclosure Avoidance Techniques, May 1978.

6. Kasprzyk, Daniel, "Social Security Number Reporting, the Uses of Administrative Records and the Multiple Frame Design in the Income Survey Development Program," Technical, Conceptual and Administrative Lessons of the Income Survey Development Program, Social Science Research Council, New York, New York.

DISCUSSION

Joseph Steinberg, Survey Design, Inc.

INTRODUCTION

The three papers presented illustrate three of a number of varying objectives of exact matching:

(1) addition of data from second file to host file for the same IRS business tax unit;
(2) construction of a more comprehensive frame by combining files; and
(3) addition of variables on establishment economic data to data for individuals in the Survey of Income and Program Participation (SIPP).

This discussion primarily comments on earlier drafts of these papers.

These papers describe the files used and how the matching was done in fine detail. I leave it to those more expert to comment on these matters; I will not try to comment on that.

PERSPECTIVE OF COMMENTS

The point of view taken in preparing these comments was:

(1) How does the quality (or likely results) of the exact matching conform to statistical standards used to judge a statistical study or to judge completeness of a frame?
(2) After reading or listening to the paper, what is known about factors (and their magnitudes) affecting the nonsampling error component of the results?
(3) What additional information should be made available to judge the nonsampling error?
(4) What more (should) might possibly be tried to reduce the nonsampling errors?
(5) Further, if a sample reinterview program is considered useful in measuring coverage and content (net and gross) differences in a sample survey or census, why not use a sample reinterview program for evaluation and calibration in matching studies?
(6) Is the matching approach optimal or is it better to collect data through a survey process?

In view of the review approach, you will see that this discussion provides some comments and a series of questions for the presenters.

GREENIA

Nick Greenia has an interesting problem, even though both files come from IRS forms. The supplementary forms for individuals (C, F, and 4835), which are of interest, may not show the EIN or, if EIN is shown, it may be incorrect. What is known (if anything) about false nonmatches or false matches as a result (since only the 1979/1980 files of the Forms 941/943 were used, and not 1978/1979)? What is known about the false nonmatch rate which resulted?

It is interesting to observe that many identifier systems have similar problems -- here it is the "sole propietorship/corporation connection" re the EIN. There used to be (and may still be) the problem in the SSN: multiple people gave an identical SSN as a result of the purchase of a wallet that had a valid SSN on a specimen identification card.

I noted that matched cases were dropped when the 941/943 payroll was greater than the sole proprietorship's business deductions. Was any effort made to contact any sole proprietorship when this was found? Wouldn't it be of interest to know, for a small sample, at least, under what circumstances this situation arose? May not treating such cases as unmatched eliminate an important class of novel situations? Why do you think, Nick, that reweighting overcomes the problem?

Given the assertion in the paper "... that a significant portion of true matches remained to be found ..." (Section V), would the analytic objectives be served if the tabulations of "matched" data are based on not much more than the original set of matches? Would the nonsampling error of the results be too large?

I have often wondered whether information on the Forms W-3 was available on any accessible file. Since the Form 941 employment is only for employees for the pay period ending March 12, would a more useful source of employment and payroll be:

(1) the number of statements--counts of Forms W-2 and
(2) total payroll for the year from the summary W-3 process?

Incidentally, if any of these questions suggest a need for contact with a business (as re 941/943 payroll greater than business deductions), a statistical study (perhaps conducted by a third party) should be considered the vehicle, with results available to IRS only in tabulations (screened for disclosure problems). Consider, a statistical reinterview program may be a useful means for evaluating overall quality and not just for special issues.

HIRSCHBERG

Now I turn to Dave Hirschberg's paper. In the paper, I found the interesting points:

(1) that the Master Establishment List (MEL) is unique in its representativeness of small businesses of all size categories, and
(2) that the total number of businesses included in the MEL exceed more than half of the population or universe of all (small and large) businesses reporting to the Internal Revenue Service.

My question is: How complete is MEL? The tables show the relation of the Duns Market Identifiers (or DMI) to County Business Patterns. How do the distributions of MEL compare with some standard? And, by Standard Industrial Classification (SIC code) and employment size?

At another point, the author indicates that businesses not represented in the MEL are mostly smaller businesses or individuals that might be located in their homes or who, due to limited activities, would not appear in the credit markets nor advertise in the yellow pages.

In view of this, what problems are there in the Small Business Administration (SBA) use of MEL? Also, what is known about the rate of inclusions in MEL files of firms no longer in existence (given the slowness of purge of the DMI and Market Data Retrieval, Inc's "yellow-page" listings)? What is the duplication rate still in the file? (One source paper says "... hopefully relatively few.") Further, what is known of the proportion of false matches -- discards from one file or the other that really didn't match? This is not to suggest that "Findit" as a match program has any discernible problems -- at least to my knowledge.

Now, I turn to another matter. This project, creation of MEL, was initiated since there was essentially no single file available to SBA which satisfied its needs--and it is understandable why various agencies have statutes (Census) or regulations which require confidentiality of frames, privacy being deemed more important than government-wide efficiency.

What is the confidentiality status of MEL? Does SBA have a regulation which prohibits disclosure? What are any other possible public uses - could another government agency, say, Department of Energy, or could a research firm doing a study for a government agency have access? At what price? How does this compare to your costs?

On another matter -- what improvements in file completeness would there be from access to the UI files in the 25 states willing to share their files? Has anyone explored the possibility that uniform files for these 25 states may exist in a Federal agency's hands -- the Bureau of Labor Statistics (BLS)? And what cooperation can be worked out between SBA and BLS, given written agreement by these 25 states to permit SBA access?

The paper recognizes that data collection is "non-rigorous" and, therefore, employment, and possibly SIC codes, too, may be inaccurate. What, if anything, can be said about the effects of possible inaccuracies on the use of subsets of MEL as survey frames? Consider the value of a sample reinterview program to check on quality.

Finally, the paper mentions that some checks were planned, e.g., MEL vs. University of Michigan, Survey Research Center's sample of their nonhousehold establishment list. Are there any results of such checks available? What do they show about the completeness of MEL?

SATER

Now concerning Doug Sater's paper; first, I turn to the SIPP information collection to be used for the match. Has Census considered the desirability of expanding the questions being asked (name of employer, address, employer identification number)? Perhaps, in addition to address (or, if not available), they could consider getting nearest street intersection; asking for telephone number at place of employment -- for possible use, when no EIN is given, for calling the employer; or, if no address, calling to establish an address?

Also, re SIPP-collected data -- what steps are taken to assure that SIPP-collected EIN is consistent with SIPP-collected information on employer name and address?

The paper discusses a prospective matching project, and it is interesting to read about the decision process that leads to the decision concerning the source file and matching method. It will be interesting to hear, in the future, what actually took place: the degree of manual effort and the various costs. Incidentally, what is the relative budget planned for this matching activity compared to the SIPP data collection phase? It would be interesting to know, both here and in other matching projects, about relative budgets for matching vs. data collection of source surveys.

In view of the author's contention that they expect to obtain (in the SIPP) valid EINs about 40 percent of the time and that there is a need to use a variety of methods to try to determine the EIN in the remainder, how will the match validity be tested? (The paper says error measurement will be the subject of future development. And evaluation strategies will be the subject of future development.) What about considering a sample reinterview program as part of the evaluation strategy?

The paper describes a small scale familiarization test. Admittedly, it was not a true test, since address and EIN had not been collected in the nonprobability set of units used for the test.

How secure are you, Doug, in the rates of exact matching cited in the paper? Do you have plans for another, truer, test, using a subsample of the SIPP that you plan to use, before mounting the full-scale matching project? Suppose the results are not as good as in the small-scale familiarization test; what if the results suggest a 60-70 percent match rate. Would you recommend the project move forward?

The paper notes that adjustments are planned for matching problems. What order of magnitude of matching problems do you believe are likely to occur, for which allocation or reweighting is the preferred solution? What do you anticipate will be the net effect on the level of nonsampling error in some principal result?

REJOINDER

Nick Greenia, Internal Revenue Service

The discussant's observations are, of course, most appreciated and exhibit a grasp of the Sole Proprietorship Link Study's fundamental problem: as a first time study, it had to cope with how much was simply unknown.

The decision to employ the 1979/1980 file of Form 941/943 records and omit the 1978/1979 file as well as the fiscal filing period possibility was due to two factors: higher processing costs and the 1979 calendar filing period assumption. Higher costs of additional linkage processing for files not originally designed for the link studies per se (i.e., the SOI-perfected sole proprietorship sample file and the Census-perfected Form 941/943 population file) were deemed unwarranted primarily because (a) for Tax Year 1979 some 99% of all Forms 1040 had calendar year 1979 filing periods and (b) of those which had fiscal or non-1979 filing periods, many were probably filed for members of partnerships.

Other than what is known of false matches obtained from match processing as well as the increase in aggregate data resulting from reweighting for false non-matches (increases of 16% for number of businesses, 10% for payroll, and 11% for employment), nothing is known of this processing decision's direct impact on false matches and non-matches. Probably it had little impact since match problems in general were thought to be attributable primarily to the Employer Identification Number (or lack of it) on the sole proprietorship's business schedule. The second Sole Proprietorship Link Study (Tax Year 1982) is expected to benefit from the 1979 experience in this regard primarily because such tradeoff decisions as necessitated for the 1979 Study will be precluded by the 1982 sample file format design.

No sole proprietorships were contacted during the study's match processing phase primarily due to resource constraints. Although the payroll/deductions discrepancy was designed to catch "hidden payroll" on the business return, the 1982 study probably will compare payroll to proxy payroll. This change is suggested by the 1979 experience which has led us to believe that hidden payroll is less of a potential problem than the overstating of proxy payroll--primarily due to its inclusion of contract labor payments as well as payroll not reportable on Form 941/943 for certain employee classes. Again though, it is important to err on the conservative side (particularly when examining the payroll/deductions relationship) by building a sound match base, due to the large weights on some sample business records. Reweighting is thought to overcome potential problems of omission by compensating for any marginal matches missed through groups of solid match records with similar characteristics. Further, it was a desirable step in order to provide the Small Business Administration (SBA) with as full a data set as possible to meet SBA's own analytic needs.

The discussant's suggestion to replace the Form 941 file with W-3 file counterpart information (total conpensation for payroll, number of W-2's attached as an employment proxy) would be desirable if control problems currently confronting the W-2/W-3 tapes--annually provided to IRS by SSA for the Combined Annual Wage Reporting Agreement Form 941/943 reconciliation effort--could be overcome. SSA is planning to overhaul its current computer processing system in 1987, which might be a more appropriate time to reconsider such an approach. In the meantime, however, it might be worthwhile to pursue this idea with the thought of supplementing Form 943 information--weakened in the past by reporting qualifications as well as the general problem of reporting employment only for the March 12 pay period.

REJOINDER

David Hirschberg, Small Business Administration

Joseph Steinberg's questions regarding the Master Establishment List's (MEL) quality and conformity to statistical standards lie at the heart of the matter, once the major issue of mechanically merging files is solved.

Limited opportunity exists here for full discussion of the quality issues raised by Joseph Steinberg. However, there are several studies and reports which provide the interested researcher with such information. Discussions of the overall quality of the Dun's DMI file can be found in "D&B, DMI: Data User Conference." [1] Another publication of interest includes, "A Comparison of Employment Data From Several Sources: County Business Patterns, UI and Brooking's USEEM," by Candee Harris. [2] That report provides a fairly extensive examination by industry of the small business population.

Generally the nonsampling errors which are of concern can be examined from the information presented. The impact of the matching on the overall quality of the MEL is more complicated. From a statistical point of view, little is known about how completely the "yellow pages" cover the universe of business.

Definitive efforts to evaluate the Master Establishment List are hampered by the lack of uniform numerical identifiers in the various systems. Even when numerical identifiers, such as Federal employer identification numbers, are available, the matching of files from different systems is not a straightforward task, as Nick Greenia has pointed out in his paper. [3]

A great deal of work is needed in this area, and access to administrative records from State and Federal agencies is necessary. In addition, a requirement exists to more carefully define a small business for statistical purposes.

The overall documentation of the Small Business Data Base work can be found in the appendices to the "State of Small Business: A Report of the President" for each year beginning with 1982. [4] A more comprehensive guide to information relating to specific issues can be found in "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography" by Bruce D. Phillips. [5] Most of these publications are available from the Office of Advocacy. Methodological and quality issues raised by Steinberg are directly addressed. Steinberg also raised the issue of the MEL's confidentiality status. This is now under discussion with the firms producing the files, and a formal statement on this issue should be forthcoming.

As mentioned previously, the inability to match files of business firms, along with a large turnover rate, plagues any attempt to develop independent verification of the MEL. The University of Michigan Survey Research Center report, although vigorous in its approach, was not able to overcome these problems. [6] When differences between the two files occurred, it was difficult to determine precisely what the problem was.

One final comment with regard to the State unemployment insurance data is in order. The potential use of these files was explored with the States and the Bureau of Labor Statistics; because of confidentiality provisions, access could not be provided. Although a few States did decide to make their files available for research purposes, the cost involved in integrating them into the MEL precluded their use.

REFERENCES

[1] Advisory Commission on Intergovernmental Relations. (1979) "D&B, DMI: Data Users Conference," Washington, D.C. (mimeographed).

[2] Harris, Candee S. (1981) "A Comparison of Employment Data for Several Business Data Sources: County Business Patterns on Unemployment Insurance, and the Brookings U.S. Establishment and Enterprise Microdata File," The Brookings Institution, Washington, D.C.

[3] Greenia, Nick (1985) "1979 Sole Proprietorship Employment and Payroll: Processing Methodology," Record Linkage Techniques--1985, Internal Revenue Service, Washington, D.C.

[4] Small Business Administration. (Annually, 1982 to 1985) The State of Small Business: A Report of the President.

[5] Phillips, Bruce D. (1985) "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography," Record Linkage Techniques--1985, Internal Revenue Service, Washington, D.C.

[6] Converse, Muriel and Heeringa, Steven G. (1984) "An Evaluation of the Accuracy and Current Utility of the 1981 Master Establishment List (MEL)," Institute for Social Research, University of Michigan, Ann Arbor, MI.