SOLE PROPRIETORSHIP EMPLOYMENT AND PAYROLL ESTIMATION: A RECORD LINKAGE APPROACH

Charles Day, Internal Revenue Service

The Internal Revenue Service, with the sponsorship of the Small Business Administration, conducts periodic studies with the goal of creating data sets containing both employment and financial data for Sole Proprietorships. While other agencies produce employment and payroll data for the same entities, none of these agencies is able to include the broad range of financial data available on the IRS files. This paper documents the methods used in constructing such a data set for Tax Year 1982.

The paper begins with a general discussion of the goals, approach, and limitations of the study. Following this, a detailed description of the linking and false link processing steps is presented, along with the results of each step. Next, an attempt is made to evaluate the coverage of the link using external control comparisons. Then, the methodology for and results of false-nonlink adjustment are described. Finally, the results of the adjustment will be briefly discussed.

This paper is a natural outgrowth of earlier record linkage work for business sector returns. Plans for this study and details of related work are provided in [1-6].

STUDY DESCRIPTION AND LIMITATIONS

The Sole Proprietorship Employment and Payroll study attempts to create a research file containing both employment data--like wages paid and number of employees--and a broad range of financial data--like receipts and costs--on sole proprietorships. The term "sole proprietorship" will be used in this paper to mean the business entity represented by a single Schedule C (Profit or (Loss) from Business or Profession) or Schedule F (Farm Income and Expenses) attached to a Form 1040 (U.S. Individual Income Tax Return) [7]. The study links an extract from the Tax Year 1982 Statistics of Income (SOI) Individual/Sole Proprietorship (INSOLE) sample file containing Schedules C and F data from Form 1040 for Tax Year 1982 to a file of Forms 941 (Employer's Quarterly Federal Tax Return) and Forms 943 (Employer's Annual Tax Return for Agricultural Employees) for Tax Years 1981, 1982, and 1983. (The data in this file have been subjected to testing and imputation for missing items by the Census Bureau.)

As might be expected, such a method introduces several potential limitations:

- First, the Statistics of Income Division does not specifically design its sample file for the accurate estimation of employment.
- Second, the Tax Year 1982 INSOLE file contains, along with returns representing sole proprietorships' 1982 Fiscal Years, other returns being filed for periods prior to the firms' 1982 Fiscal Years, which were processed during Internal Revenue Service (IRS) Processing Year 1983. Some of these records had to be eliminated from consideration for linking, as they were filed for accounting periods for which we had no Form 941 data. (These were later considered nonlinked returns for purposes of false nonlink adjustment.)
- Third, this study uses the Employer Identification Number (EIN) as the linking variable. It is important to note that the Business and Individual Master Files are distinct entities. (The former provides the employee information, the latter contains the financial data.) The number used to identify an account on the Business Master File (BMF) is the EIN; the number used to identify an account on the Individual Master File (IMF) is the social security number (SSN). Theoretically, this does not present a problem, since proprietors who have employees and file Forms 941 and 943 (the target of this linkage) are required to have EINs [8], and are instructed to supply them on their Schedules C and F. In practice, however, since not all sole proprietorships are required to have EINs [9], and the EIN is not necessary to the revenue collection function, there is little done to detect and remedy missing EINs. More information regarding the rate and causes of EIN absence is contained in a longer version of this paper [10]. This "missing EIN" syndrome is one cause of a further difficulty which plagues all exact match studies--false nonlinks. Both this problem and its companion--false link--were present in this study. Efforts were made to address both of these phenomena, as discussed below.
- Finally, data on farm employment (from Form 943) were found to be too unreliable for use, as no effort is made by either IRS or Census to detect even obvious errors, much less more subtle ones. For this reason, farm employment data are limited to those farm returns which matched a Form 941 record, and they do not, in any way, represent a comprehensive measure of farm employment.

PROCESSING METHODOLOGY

Description of the Data Files

The Forms 941/943 file consists of one record for each EIN, containing employment (reported annually for the pay period including March 12); payroll from Form 941 (quarterly data) or Form 943 (annual data); a Census-derived industry code [11] for each year; and a set of filing requirement codes extracted from the entity portion of the IRS BMF [12]. The "raw material"

from which Census created this file was an extract of BMF data which had multiple Form 941 filings for a single EIN consolidated into a single record.

The Sole Proprietorship file was created by extracting the Schedules C and F information from the Tax Year 1982 Individual/Sole Proprietorship file. The INSOLE file contains records representing a stratified probability sample of the Forms 1040 filed for Tax Year 1982 [13]. Population estimates are made from this file, using weights based on a return's likelihood of being sampled. Each variable-length INSOLE file record contained, at most, three Schedules C and two Schedules F [14]. Each record in the extract file represented one Schedule C or F.

Link Definition

Probably the most critical element in any record linkage is the definition of a true link. The EIN is a strong linking variable; this allowed the adoption of a relatively simple linkage rule. Two records whose EINs matched on all nine digits were initially defined as a true link; strong and convincing evidence was required that the records represented different reporting units before they were redesignated as falsely linked [15].

Matching Errors

Two types of errors arise in the matching process. The first is a false link; that is, two records, one from the Schedules C/F file, the other from the Forms 941/943 file, which link on EIN, but represent different reporting units. The second is a false nonlink, where records which should link don't because of problems with EINs. Each of these errors and its subsequent adjustment is discussed in the sections that follow.

FALSE LINK ERRORS AND ADJUSTMENTS

False Links

First consider links between records representing different reporting units. Often, such mismatches between two records due to an error in the linking variable (EIN) may be detected by comparing variables which have a well-established relationship between the records. In order to ensure against false links, the following test was employed in this study: Fiscal Payroll (line 2, Form 941 or line 2, Form 943) was compared to Total Deductions from Schedule C or F. There is no guarantee that the explicit payroll deductions on the sole proprietorship schedule will sum to Fiscal Payroll for the period of the return, since some direct payments to workers could be hidden in other deduction items, such as Repairs, or the catchall category, Other Deductions. It is certain, however, that the true Fiscal Payroll amount cannot exceed Total Deductions. Further, in practice, the proprietor has every incentive to minimize Fiscal Payroll, as this reduces his payroll tax liability, and to maximize Total Deductions, as this reduces his income, and, thus, his income tax liability. Therefore, links in which the Fiscal Payroll exceeded the Total Deductions on the Schedule C or F were deemed to be false links and were treated as such.

Next, in order not to inflate the financial data for linked records with data from linkages which showed no Forms 941/943 payroll or employment activity for the period covered by the Schedule C or F record, linked records with both Fiscal Payroll and Employment equal to zero were designated falsely linked.

The final complication addressed during false link processing was the occurrence of multiple Schedule C or F records linked to the same Form 941/943 record (a duplicate linkage). Note that, while this is possible due to the occurrence of multiple Schedule C or F records with the same EIN (usually representing more than one business owned by the same proprietor), the reverse (that is, multiple Forms 941/943 records linked to the same Schedule C or F record) is not possible, since EINs are unique on the Forms 941/943 file. The easiest cases to resolve were those sets of duplicate linkages in which one or more of the Schedule C or F records had zero Proxy Payroll (the sum of Salaries and Wages and Cost of Labor variables from the Schedule C record). The absence of Proxy Payroll was taken as a strong indicator that the record did not represent a business which had employees; thus, the 471 such cases were classified as false links. A manual review of the remaining duplicate cases was performed, taking into account such factors as the industry codes of the Schedule C or F record, the industry code from the Form 941 or 943 record, Proxy Payroll, Fiscal Payroll, BMF filing requirement code (used to detect partnership, corporation, or nonprofit organization Forms 941 or 943 which linked to sole proprietorship records), Total Receipts, and Total Deductions. This review resulted in the classification of another 153 records, with 0.7% of Proxy Payroll, as false links, and the classification of 1,002 records, containing 8.6% of Proxy Payroll, as true links. In the true link cases, the Form 941 or 943 was deemed to represent the one or more remaining records in the group, and Fiscal Payroll and Employment were apportioned among them based on each record's proportion of the EIN-group's total Proxy Payroll. A summary of link statuses, counts of records, and amounts of Proxy Payroll are available in columns one and two of Table 1.

Evaluation of Link Coverage

Table 2 compares the initial results of the Sole Proprietorship Link with three other estimates of Sole Proprietorship employment. First, data from the Census Bureau's Enterprise Statistics program for 1982 were compared to the link studies results. Despite the limitations on comparison of data from these sources [16], the need for adjustment of the Employment and Payroll Link data is apparent. The percentage of the Census estimate of employment represented by the unadjusted Link estimate ranged from a low of 57.3% for Mining to a high of 88.9% for Construction. For the important Retail Trade and Services industries these percentages were 65.4% and 69.9%, respectively.

A second source of comparison data was the Tax Year 1979 study. While some growth and

TABLE 1. NUMBER OF RECORDS AND PROXY PAYROLL BY MATCH STATUS

Match Status Description	Original Weights		Adjusted Weights	
	Number of Records	Proxy Payroll	Number of Records	Proxy Payroll
Frequencies and Amounts				
Total	13,885,209	47,123,960,894	13,221,235	44,921,342,868
Transcribed duplicates	7,118	227,164,875	2,215	30,597,051
Duplicates with Proxy Payroll = 0	18,759	0	18,607	0
Total Deductions less than Fiscal Payroll	89,561	134,863,801	82,669	11,252,701
Unlinked records w/ EINs	585,436	2,332,971,146	475,655	315,264,291
True links	988,747	26,095,651,037	1,433,492	37,969,720,316
True linked duplicates	39,693	1,163,385,998	57,005	1,655,398,980
Out-of-Scope	18,107	442,086,925	7,055	30,616,080
Zero Fiscal Data	126,317	487,726,041	96,944	93,365,990
No EIN	12,011,435	16,240,111,071	11,047,593	4,815,127,459

TABLE 2. COMPARISON OF LINKAGE RESULTS WITH EXTERNAL ESTIMATES

Industry	1982 Link		1982 Census*		1979 Link		1979 Adjusted Link&	
	Number of firms	Number of employees	Number of firms	Number of employees	Number of firms	Number of employees	Number of firms	Number of employees
Mining	2,429	12,676	5,000	22,113	5,166	26,791	7,355	38,143
Construction	115,980	345,610	122,901	388,700	213,440	583,274	234,784	641,601
Manufacturing	36,102	136,480	45,357	215,468	51,257	239,777	54,332	254,162
Transportation, Communication, and Public Utils.	33,474	131,266	**	**	47,950	172,353	43,991	158,123
Wholesale Trade	27,618	116,580	47,313	165,473	43,715	154,589	38,469	136,038
Retail Trade	303,741	1,201,360	446,774	1,835,761	512,873	1,933,676	548,774	2,069,033
Finance and Insurance	24,932	45,656	**	**	32,432	62,980	&&	&&
Real Estate	10,417	25,931	**	**	29,211	48,821	&&	&&
Services	368,754	1,035,584	510,077	1,481,530	518,726	1,392,895	606,909	1,629,686

NOTE: The Sole Proprietorship Employment and Payroll Study does not attempt to estimate employment for Farming; therefore, this table omits the Agriculture, Forestry, and Fishing division.

* These estimates are taken from the Census Bureau's 1982 Enterprise Statistics series [17].

** These industries are not covered in the Enterprise Statistics series.

& These are 1979 linked data crudely adjusted by taking the ratio of the total (SOI) number of firms in each industry in 1982 to the number for 1979, and applying that ratio to the 1979 employment link data as a weighting factor.

&& These data were not listed separately in the available Sole Proprietorship tables.

distributional change might be expected, one would also expect that these estimates and the Tax Year 1982 estimates would fall into the same "ballpark." The unadjusted 1982 data do not meet these expectations. In fact, the percentage of the Tax Year 1979 employment data represented by the Tax Year 1982 employment data ranges from a low of 47.3% for Mining to a high of 76.1% for Transportation, Communication, and Utilities. Again, the Retail Trade and Services industries perform disappointingly, showing 62.1% and 74.4% of their Tax Year 1979 values, respectively.

Finally, comparisons were made after a crude adjustment was applied to the 1979 data. For each major industrial division, the number of sole proprietorships with employment in the 1979 employment and payroll study was multiplied by the ratio of the total number of sole proprietorships in that industry in 1982 to the total number of sole proprietorships in that industry in 1979. That is, the results of the 1979 study were adjusted using the growth rate of all sole proprietorships between 1979 and 1982. This provided a rough adjustment for distributional changes and growth. Ratios between the unadjusted 1982 data and the adjusted 1979 data ranged from 33.2% for Mining to 85.6% for Wholesale. The comparison for Retail was 58.1%, while it was 63.5% for Services.

FALSE NONLINKS AND ADJUSTMENTS

Problem Definitions

All three sets of comparison estimates confirmed the need for adjustments in the 1982 link data. There remained one problem likely to have had an effect on the data of the magnitude and direction observed; that is, the presence of false nonlinks which occur when a record in the Sole Proprietorship file and a record in the Forms 941/943 file represent the same reporting unit, but, due to an imperfection in the linking variable, or a missing linking variable, they do not link. The causes of this phenomenon will be discussed, then the procedures undertaken to adjust for the errors will be described.

False Nonlinks

At least two well-defined causes for false nonlinks exist within this study file. The first cause is a missing EIN; this problem has been discussed previously, but it is worth reiterating here that records with no EIN occur due to a process different from that which causes records with erroneous EINs. It is also worth noting again that most records without an EIN are not in error.

The second cause of false nonlinks is an erroneous EIN. This can occur due either to taxpayer misreporting, or to transcription errors during data entry. Note that these two sources of error are not differentiable. Furthermore, it is possible for a sole proprietor who legitimately obtained an EIN for past needs (e.g., registration of a Keogh plan, or filing an employment tax return) to report an EIN on Schedule C or F without being required to file Forms 941 or 943. The two causes of false nonlinks imply the need for a two-part correction--first, to adjust for the records without EINs which should have had EINs, and, second, to estimate payroll and employment for those records with EINs (now including those adjusted for in the first part of the procedure) which should have linked, but did not.

Adjustment Methodology

Description.--Adjustment for false nonlinks due to erroneous EINs (henceforth referred to simply as false nonlinks) and false nonlinks due to absence of EIN (henceforth referred to as false absence of EIN) employed similar techniques. A 13 x 12 x 10 x 11 x 3 analytical table, in which each cell contained a count of records with the characteristics associated with that cell, was created. (The first four dimensions correspond to stratifications of Industry, Size of Adjusted Gross Income (from the associated Form 1040), Size of Business Receipts, and Size of Proxy Payroll. The last dimension consists of Linked, Unlinked with EIN, and Unlinked without EIN categories. Additional details about this table are given in [10].

The false absence of EIN adjustment procedure began by assuming that some region of the analytical table could be defined, within which 100% of the records should have had EINs. Within each cell of that region, an adjustment factor was computed equal to the inverse of the rate of occurrence of EINs. Within each major industrial division, the weighted median of these factors was applied to all of the cells outside the 100%-EIN region, on the assumption that the process creating false absence of EIN was random with respect to the other dimensions of the table. Taking the difference between the original value in the With-EIN subtable and the results of applying this factor to the original cell value resulted in an adjustment amount; that is, a number of records which needed to be added to the With-EIN cell value, and subtracted from the corresponding cell value in the Without-EIN subtable in order to properly correct for false absence of EIN. These additions and subtractions were then performed.

This adjustment had the effect of increasing the Unmatched-with-EIN category to account for the records whose EINs were falsely absent; the result was that, for purposes of adjustment for false nonlinks, the missing EIN problem could be considered corrected. Then, reweighting to adjust for overall false nonlinks from all causes could be undertaken in a single procedure, as follows. Let

$$X_{A(\text{linked})} = \sum_{i=1}^{M} X_{Ai}w_i$$

$$X_{A(\text{false nonlink})} = \sum_{i=M+1}^{M+N_F} X_{Ai}w_i$$

$$X_{A(\text{true nonlink})} = \sum_{i=M+N_F+1}^{M+N_F+N_T} X_{Ai}w_i$$

$$X_{A(\text{total})} = \sum_{i=1}^{M} X_{Ai}w_i + \sum_{i=M+1}^{M+N_F} X_{Ai}w_i + \sum_{i=M+N_F+1}^{M+N_F+N_T} X_{Ai}w_i,$$

where the file is conceptually ordered in such a way that the first M records represent true links, the next N_F records false nonlinks, and, finally, the last N_T records true nonlinks; X_{Ai}, denotes the sampled value of Ath item in the ith record; and w_i represents the weight determined by the rate at which returns in the record's class were sampled in the 1982 SOI Sole Proprietorship study.

The aim of the reweighting is, then, to develop a set of unit reweighting factors (F_1, F_2, ..., F_m), such that

$$\sum_{i=1}^{M} F_i(X_{Ai}w_i) = \sum_{i=1}^{M} X_{Ai}w_i + \sum_{i=M+1}^{M+N_F} X_{Ai}w_i$$

$$\sum_{i=M+1}^{M+N_F+N_T} F_i(X_{Ai}w_i) = \sum_{i=M+N_F+1}^{M+N_F+N_T} X_{Ai}w_i.$$

Implementation

Data Reduction.--The first stage in implementing the adjustment methodology involved reducing the analytical table. The original table contained 51,480 cells, most of which contained zero values. This table may be usefully thought of as consisting of three subtables. One contained the (weighted) number of Schedule C records which linked (these necessarily had EINs) within each combination of descriptive categories defined by the Business Receipts size, Adjusted Gross Income size, Proxy Payroll size, and Industry dimensions; a second contained the number of such records which had EINs but failed to link; a third contained the number of such records which did not have EINs (these necessarily did not link). The stratifications represented by the four dimensions of each of these subtables were arrived at a priori, and represented the best guess as to what categories would be analytically useful. The table which resulted

from these assumptions turned out, in fact, to be large and sparse. When contingency table analysis was attempted on this table, spurious results were obtained. In order to collapse the table so that it would be less sparse, the table strata were collapsed in such a way that no stratum contained fewer than 50 nonzero cells. Next, the table strata were tested for their effect on the distribution of the data in the table by using a computer routine called EFFECTS. This routine used a logit model to quantify the effect of a record's being in a given category of one of the dimensions on the likelihood of lying within each of the subtables. The goal of this collapsing was to reduce the proportion of zero cells to a minimum constrained by the desire not to eliminate analytically useful categories. The result of this work was a 10 x 7 x 7 x 6 x 3 table.

Following this initial collapsing, contingency table analysis was used to reduce the size of the subtables by examining the predictive value of each of the dimensions with respect to the distribution of the data within the table. A computer routine called CONTAB was used to construct alternative tables to the analytical table under the assumption of simpler interactions between the predictive variables (the subtable dimensions) and the dependent dimension, containing the EIN-presence and link status categories (the categories which divide the original analytical table into subtables). Tables were created assuming many different models, and compared to the original table. More specifically, a measure of the distance between two tables based on the Minimum Discrimination Information Statistic was employed. A set of constraints, in this case sets of marginal totals of the original table which must be maintained, was specified. A fitted table was created in such a way as to minimize the value of the following formula, subject to the set of constraints:

Let I = Minimum Discrimination Information Statistic
P = A value from the fitted table
ρ = A value from the original table

$$I = \sum_{ij} P_{ij} \ln (P_{ij}/\rho_{ij}).$$

In implementing this procedure, we proceeded by creating two tables, the first was based on a set of constraints corresponding to a model which allowed all of the two-way interactions of link status and the predictive variables. The second was created by omitting one set of constraints, representing the interaction of one of the predictive variables with link status, from the first model. By using models allowing only two-way interactions between predictive variables and link-status variables, it was possible to identify the interaction of each predictive variable with link status. The information number of the second model was compared to that of the first in order to measure the predictive power of the omitted variable. This was done one at a time for each of the predictive variables. In so doing, it was discovered that AGI had little, if any, predictive power, and could be removed from consideration. The results of this procedure led to the collapsing of the 5-way table into a 4-way, 10 x 7 x 6 x 3 table, exclusive of AGI. The results of the contingency table analysis, and the categorical dimensions of the new table are described in [10].

Adjustment for False Absence of EIN.--After collapsing the original table as described above, the adjustment procedure moved on to the next step, development of a correction for false absence of EIN. As described in the previous section, this involved one more collapsing step, addition of the linked and unlinked records with EIN, to form a new 10 x 7 x 6 x 2 table, the last dimension of which contained the categories With-EIN and Without-EIN. A region of this table, generally corresponding to areas in which both Business Receipts and Proxy Payroll were high, was identified in which one could assume 100% of the records ought to be reporting payroll taxes, and, thus, should have an EIN. Both qualitative assumptions about the likelihood of firms with high sales or large labor compensation deductions having employees, and the behavior of the rates of presence of EIN were used to identify these regions. The inverse of the rate of presence of EIN was then computed for each cell of this region, and, within each major industrial division, the weighted median of these ratios was used to compute an adjustment for the non-100% region of the table. The appropriate amounts were then added to the Unlinked-with-EIN cells of the original table, and subtracted from the Without-EIN cells to effect the adjustment.

False Nonlink Adjustment.--Following this adjustment to the original table, a new 10 x 7 x 6 x 2 version was produced, temporarily ignoring the Without-EIN adjustment. Note that this was necessary, as that adjustment would confound the observed link rates and make definition of a 100%-link region difficult. The link-nonlink table had "Linked" and "Unlinked-with-EIN" as the categories of its last dimension. Once again, in regions where the amounts of Business Receipts and Proxy Payroll were both high and where the observed link rate was also high, the assumption that 100 percent of the records should have linked was adopted. This process identified one set of records--those for which the strong assumption of 100 percent linking could be made. The "gray area" records, the false nonlinks which did not fall into this category were addressed by assuming that the same false nonlink process operated outside the 100 percent link region as inside it. Thus, a decision was made to adopt the weighted median adjustment factor for the 100 percent region of a particular industry as an adjustment factor for the rest of the records in that industry.

Applying a method similar to that used for the false absence of EIN, reweighting factors were then developed for the 100 percent region and applied. The factor applied to the linked records in each cell in the 100 percent region of a given industry consisted of the inverse of the link rate for that cell. Following this, an overall factor for the linked records in the non-100 percent link cells, equal to the

weighted median adjustment factor for the 100 percent region cells, was calculated on an industry-by-industry basis. Finally, an effective factor, equal to the minimum of the overall factor or the factor which caused the number of adjusted linked records to equal the sum of original linked plus nonlinked records, was produced for each cell.

Next, a set of factors for the nonlinked records was calculated, on a cell-by-cell basis, such that the sum of the linked and nonlinked records in each cell was held constant after application of the adjustment factors to both the linked and unlinked records. The application of these factors to the file resulted in adjustment of the file for false nonlinks. After this reweighting, the file was considered final; false links had been removed, and both false nonlinks due to erroneous EINs and to missing EINs had been adjusted for by reweighting.

Results

After false nonlink adjustment, the true link categories went from having only 57.9% of the Proxy Payroll to containing 88.2 percent. The "Without-EIN" category fell from having 34.5 percent of Proxy Payroll to having 10.7%. That left the residual nonlink categories with only 1.1% of Proxy Payroll. Much of the unaccounted for Proxy Payroll may represent contract labor. Better yet is the improvement in the critical Retail Trade and Services industries. Employment was originally underestimated in these industries by a factor of about one third. After adjustment, this study's employment estimate was 92.5% of the Enterprise Statistics estimate in Retail Trade, and 99.9% of the Enterprise Statistics level in Services.

CONCLUSION

While space constraints do not permit discussion of the evaluation efforts employed to examine the effectiveness of these adjustments, suffice it to say that they were considered to be successful. In fact, on the whole, we are quite optimistic about this approach. The full paper [10] provides further discussion of the entire project, including some results and comparisons to other internal and external sources.

ACKNOWLEDGMENTS

The author would like to thank H. Lock Oh for lending his statistical expertise to this project, Nick Greenia for a thorough technical review of this paper, and Beth Kilss and Wendy Alvey for their editorial assistance.

REFERENCES AND NOTES

[1] Rose, Paul; Taylor, Linda. (1982) "Size of Employment in SOI: A New Classifier," 1982 American Statistical Association Proceedings, Section on Survey Research Methods, pp. 298-302.

[2] Greenia, Nick. (1985) "1979 Sole Proprietorship Employment and Payroll: Processing Methodology," Record Linkage Techniques--1985. Internal Revenue Service, pp.285-289.

[3] Hirschberg, David; Phillips, Bruce. (1982) "Using Financial Data to Evaluate the Status of Small Business," 1982 American Statistical Association Proceedings, Section on Survey Research Methods, pp. 449-451.

[4] Greenia, Nick. (1985) "Partnership Employment and Payroll," 1978-1982 Partnership Returns, Internal Revenue Service, pp. 221-236.

[5] Moglen, Gail; Day, Charles and Petska, Tom. (1987) "Record Linkage and Imputation Strategies in the 1982 Business Employment and Payroll Studies," Statistics of Income and Related Administrative Record Research: 1986-1987, pp. 145-153.

[6] Petska, Thomas. (1985) "Studies of the U.S. Business Sector through Microdata Record Linkages," Multinational Tax Modelling Symposium Proceedings, Revenue Canada Taxation, pp. V-11 through V-20.

[7] This definition differs from the one used in the Statistics of Income Tax Year 1982 Sole Proprietorship study; that study defined a sole proprietorship as any Form 1040 with one or more Schedules C or F attached. For more information, see Wolfe, Raymond, "Sole Proprietorship Returns, 1982," Statistics of Income Bulletin, Vol. 4, No. 1, Internal Revenue Service, Washington, DC, Summer 1984, pp. 17-44.

[8] Since Forms 941 and 943 must be posted to the BMF, and an EIN is used to identify BMF accounts, then Forms 941 and 943 must have EINs.

[9] Form 1040, to which the Schedules C and F are attached, posts to the Individual Master File on the proprietor's Social Security Number; other estimates of the number of Sole Proprietorship employers show that they are a minority of Schedule C and F filers, representing 10-15% of the schedules filed.

[10] Day, Charles, "Sole Proprietorship Employment and Payroll Estimation: A Record Linkage Approach." Statistics of Income and Related Administrative Record Research: 1988-1989, forthcoming.

BIBLIOGRAPHY

Scheuren, F. J. (1973) "Ransacking CPS Tabulations: Applications of the Log Linear Model to Poverty Statistics," Am. Econ. Soc. Measurement, Volume 2, pp. 159-182.

Gokhale, D. V, and Kullback, Solomon. (1978) The Information in Contingency Tables, Marcel Dekker, New York.