RECORD LINKAGE AND IMPUTATION STRATEGIES IN THE 1982 BUSINESS EMPLOYMENT AND PAYROLL STUDIES

Gail Moglen, Charles Day, and Tom Petska, Internal Revenue Service

The Statistics of Income Division of the Internal Revenue Service (IRS), as part of a contractual agreement with the Small Business Administration (SBA), conducts periodic studies linking payroll tax returns to business tax returns for the purpose of developing a data base which includes both financial and employment data. The IRS samples of corporation, partnership, and sole proprietorship returns are a rich source of data on business activity, containing detailed income statement data for all three types of business organizations and balance sheet data for partnerships and corporations. These files, however, contain only two of the three frequently-used measures of the size of a business--receipts and assets. They do not contain the third measure, employment, because this is not reported on business income tax returns. However, employment and payroll are reported on the employment tax returns filed by these same businesses. Thus, by linking these two sets of records for the same entities, a more complete picture of business size is obtained. To date, the Corporation employment and payroll link studies have been conducted for Tax Years 1979 and 1982. In addition to providing data of interest to IRS, these studies aid in the development of the Small Business Data Base in partial fulfillment of SBA's Congressional mandate to evaluate public policy and economic trends as they effect small businesses. Because tax return information is used, the Congressional mandate can be met without placing any additional data collection burden on small businesses. Several reports on this work have already appeared in print [1,2,3,4].

The current paper will look at a number of different aspects of the Tax Year 1982 IRS Corporation employment and payroll link studies. Organizationally, the paper is divided into six parts. The first part describes the sources of data used in the sample. The second part provides an overview of the linking methodology for the Sole Proprietorship and Partnership studies. Next there is a detailed discussion of the Corporation linking methodology. This is followed by a description of imputation and reweighting for partial links and false nonlinks. The fifth part of the paper presents analysis of preliminary 1982 payroll data. Finally, there is a discussion of proposed enhancements and research activities for future business employment and payroll link studies at IRS.

SOURCES OF THE DATA

The income tax return files used in these studies are the IRS' Statistics of Income (SOI) samples of corporations, partnerships and sole proprietorships for Tax Year 1982. The figures for employment and payroll added to these files are reported by the taxpayer on the Employer's Quarterly Federal Tax Return, Form 941 series, and Employer's Annual Return for Agricultural Employees, Form 943. (Figure 1 provides a summary of each of these sources.)

Figure 1.--Sources of Data for Employment and Payroll Link Studies

Data Sources	Population	Sample Size
Corporations	2,925,933	93,675
Partnerships	1,514,212	33,557
Sole Proprietorships	13,885,209	52,020

SOI Corporation Sample

The U.S. Corporation Income Tax Return, Form 1120, reports income, gains, losses, deductions, and credits of U.S. corporations. The return is filed by domestic corporations, real estate investment trusts, regulated investment companies, insurance companies, and foreign corporations doing business in the U.S. About 94,000 corporation returns were selected from a population of approximately 2.9 million returns (Form 1120 series) filed with accounting periods beginning as early as January 1981 and ending no later than December 1983. The sample was a stratified probability sample, selected at rates proportional to size, as measured by the higher of total assets or net income/deficit for broad industrial classifications. The sample was designed to include all corporations with $10 million or more in total assets, except for corporations in the financial industries, where a minimum of $25 million in total assets was required to assure selection [5]. Approximately 38 percent of the sample returns were filed for the calendar year 1982. These included returns of most of the larger corporations. Approximately 79 percent of total assets, 63 percent of net income (less deficit), and 61 percent of total receipts were reported on 1982 calendar year returns. In addition to returns with accounting periods that spanned 12 months, the total number of active corporations includes returns with accounting periods of shorter durations. Such returns are referred to as part-year returns and were filed, for the most part, by corporations changing their accounting periods, new corporations in existence less than 12 months, merging corporations, and liquidating corporations [6].

SOI Partnership Sample

The U.S. Partnership Return of Income, Form 1065 is an information return used to report the income, deductions, credits, gains, and losses from the operation of a partnership. The return is filed by every partnership engaged in a trade

or business or having income from sources within the United States. About 34,000 partnership returns were selected from a population of approximately 1.6 million returns (Forms 1065) filed during 1983. Over 97 percent of these returns had accounting periods for the calendar year ending in December 1982. The sample was a stratified probability sample selected at rates proportional to size, as measured by the higher of gross receipts or total income/deficit, and total assets. Separate sampling rates were designed for real estate operators as opposed to other partnerships. The sample includes all partnerships in which there were $5 million or more in gross receipts, total income or deficit or total assets [7].

Sole Proprietorship Sample

While Corporation and Partnership income is reported on returns separate from those of their owners, Sole Proprietorship income is reported on schedules attached to the proprietor's individual tax return. The 1982 Sole Proprietorship estimates are based on a sample of individual income tax returns, Forms 1040, processed by the IRS during 1983. The sample was stratified based on the presence or absence of Schedule C, Profit (or Loss) from Business or Profession; Schedule F, Profit (or Loss) from Farm; the larger of total income or total loss; and the size of business plus farm receipts. Farm investors or landlords not materially participating in the business from which they received rent, and businesses operated by trusts or estates were excluded from the sample. The returns were selected at rates that ranged from 0.02 percent to 100 percent. For 1982, there were 52,020 Forms 1040 with at least one Schedule C or F attached in the sample drawn from approximately 12,000,000 such returns. These returns contained a total of 52,391 Schedules C and 16,165 Schedules F [8].

Employment Tax Returns

Employment tax returns, Forms 941 series and Form 943, provide for the reporting by employers of withheld income taxes and FICA (Social Security) taxes. They are filed by employers of all types, including partnerships, corporations, and sole proprietors. The Form 941 is filed by nonfarm employers and covers a calendar quarter. The Form 943 is an annual return used to report agricultural employment and covers four calendar quarters. In both cases, the taxpayer is required to report the number of employees on the payroll for the week including March 12 of each calendar year.

Nonfarm employers report their payroll as "Total wages and tips subject to withholding, plus other compensation" on Form 941. Agricultural employers report their payroll as "Taxable cash wages paid during the year" on Form 943. These are the figures that are added to the linked files for the current studies.

Employment tax returns are filed by the employer/taxpayer at the ten IRS Service Centers. As with other returns, various checks are made on the validity of items reported on these returns to ensure that the tax data are accurate. With respect to the processing of employment and total payroll, however, the checks are limited. This is because the withholding rates, as a percentage of payroll, vary with each employee, making it impossible to relate, with certainty, total withheld taxes to total payroll at the reporting unit level. Because of this, the IRS files of employment tax returns contain some missing and inaccurate data on employment and payrolls. Hence, it was decided to look for an alternative source for these data.

The Bureau of the Census receives Forms 941/943 tape files from IRS on a regular basis. The amounts of payroll and employment on these files are subject to additional validation and imputation for missing and incorrect data. Because of the extent to which Census processes and checks these two items, it was decided that the employment tax return data to be used in the IRS link studies should be that which Census produces. Therefore, arrangements were made to obtain the employment tax return data from the Census Bureau for all Forms 941 and 943 filed for reporting periods in calendar years 1981, 1982, and 1983.

AN OVERVIEW OF LINK PROCESSING

Defining a Linkage

The linking of employment returns to business returns is carried out on a record-by-record basis using the Employer Identification Number (EIN) as the linking variable. Probably the most critical element in any record linkage is defining a true link. A typology of linking outcomes is provided in Figure 2. Fortunately, the EIN, as a strong linking variable, allowed for adoption of simple linkage rules. When two records linked on EIN, strong and convincing evidence was required that the records represented different reporting units before they were designated falsely linked. In fact, this was only done when the linked records

Figure 2.--Possible Link Outcomes

TRUE LINK = A link between an Income tax return record and an Employment tax return record representing the same reporting unit.

TRUE NONLINK = An Income tax return record which fails to link, and for which no Employment tax return record exists representing the same reporting unit.

FALSE LINK = A link between two records representing different reporting units.

FALSE NONLINK PROPER = An Income tax return record which fails to link, and for which an Employment tax return record exists representing the same reporting unit.

PARTIAL LINK = A consolidated Corporation record for which at least one member of the consolidated group linked, at least one other member of the group failed to link, and for which there exists at least one Employment tax return record representing the same reporting unit as one of the members which failed to link.

failed a test based on accounting identities within the tax return record. While, on the whole, the EIN is a very good identifier, the following factors complicated the linkage of records and tabulation of data from linked records:

- transcription or reporting errors in the EIN;
- different reporting periods of the tax returns (e.g., a calendar quarter or year or a fiscal year); and
- noncomparability of reporting units (e.g., nonconsolidated versus consolidated returns).

Defining the Population

It was mentioned earlier that income tax returns were filed for annual, often noncalendar, periods, while the employment returns were filed for calendar quarters. This difference was handled by accumulating the appropriate quarters or fractions of quarters from the employment return data to match the linked corporation's fiscal year. EIN-linked record pairs for which no nonzero Forms 941/943 data existed for the accounting period of the income tax return were excluded from true link status. Thus, we excluded Forms 941/943 records with no nonzero data for the accounting period of the corresponding income tax return from the range of our linking function. We also restricted our linking domain. The sample files contained records which represented prior-years' returns acting as proxies for this year's late filers; if an EIN-linked pair had an accounting period which extended beyond the period for which Forms 941/943 data were available, these links were also excluded from the true link category.

Partnership Link Processing

The Partnership link processing is, on its face, the most straightforward, since only one EIN is associated with the relevant tax and payroll records, and legitimate duplication of the EIN in the Form 1065 file is limited. In a typical link for partnerships, the Form 1065 is linked to any Forms 941 or 943 with the same EIN. A new record is then created and is analyzed. The new record provides a second source of payroll and a primary source of employment information. Amounts for Fiscal Payroll, which is payroll reported on the Forms 941/943, are compared for reasonableness with amounts for Proxy Payroll (the sum of Salaries and Wages, Cost of Labor, and Rental Salaries and Wages) which comes from the partnership returns. This is done to detect false links. Also, listings of selected variables are reviewed for those cases in which an EIN was duplicated in the Form 1065 file, in order to eliminate any double-counting of Form 941/943 information due to duplicate links. For the Sole Proprietorship and Corporation studies, the processing is complicated by the possibility of multiple records of the same entity in either the tax return or payroll return files.

Sole Proprietorship Link Processing

The Sole Proprietorship linkage is quite similar in many respects to that for Partnerships. However, the absence of an EIN on most of the Sole Proprietorship records (only 20.8 percent of the 52,020 Sole Proprietorship records contained an EIN) created special difficulties in the study. Most of the returns which reported no EIN did so legitimately; Sole Proprietors without common-law or statutory employees are not required to have EINs. (Partnerships and Corporations, on the other hand, must have EINs, as these also serve as their Taxpayer Identification Numbers.) However, a significant number of returns without EINs report Salaries and Wages or Cost of Labor in excess of what can reasonably be considered incidental. This implies that the firms involved were likely Form 941 or 943 filers, but the records had no chance to link, due to missing Sole Proprietorship EINs. Such a possibility results from the fact that Sole Proprietorship employment data were processed by IRS and posted by EIN to the IRS Business Master File while Sole Proprietorship tax records, Schedules C and F, were processed with the appropriate individual income tax return (Forms 1040) and posted to the IRS Individual Master File (IMF) by that Form 1040's Social Security Number (SSN). Since the revenue processing function of IRS does not use the Sole Proprietor's EIN, little testing or perfection was performed for this variable. This created the potential for false links as well as false nonlinks. Further, legitimate duplication of EINs is much more common among Sole Proprietorship records--more than 11 percent of the records with EINs shared an EIN with at least one other record. Hence, the review required to link Form 941 information to the proper sole proprietorship return was extensive.

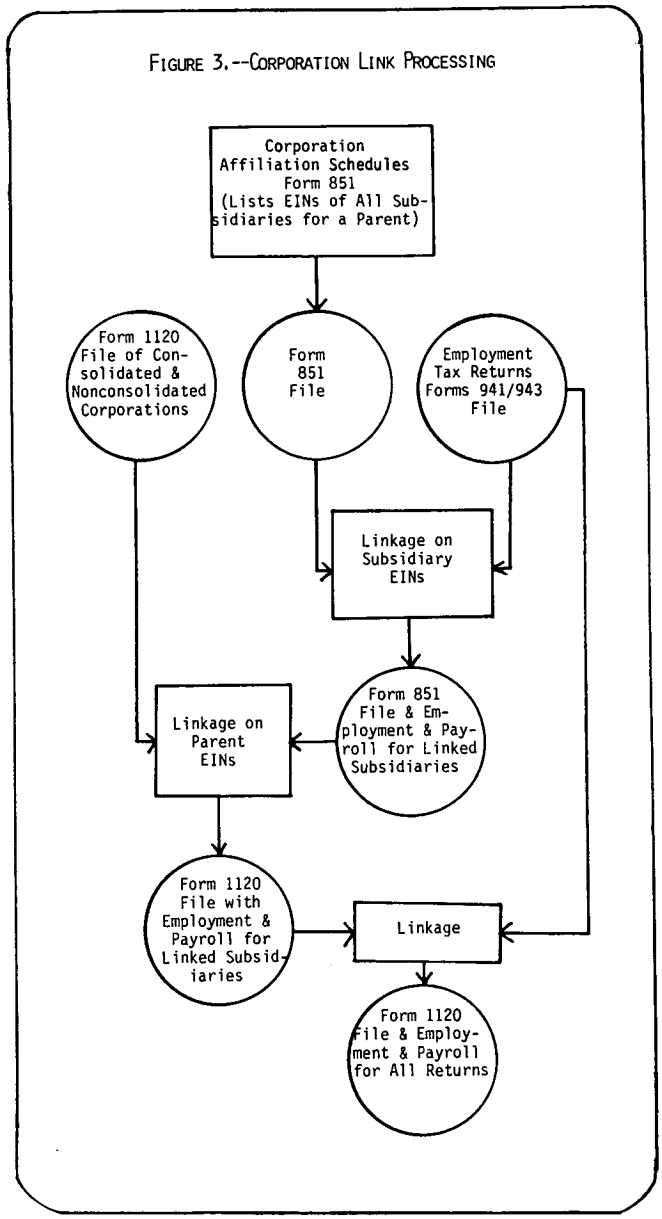
CORPORATION LINKING METHODOLOGY

The Corporation Employment Link study is more complex than the others since the SOI Corporation sample file contains both consolidated and nonconsolidated corporation income tax returns. Consolidated returns are filed by corporations which own another corporation. (See Appendix I.) These returns contain combined data for the owning (parent) corporation and the owned (subsidiary) corporation(s). Together the parent and subsidiaries form a consolidated group. The consolidated returns represent a special problem, in that a single income tax return is filed for the consolidated group, but separate employment and payroll tax returns are often filed by each of the subsidiaries as well as the parent corporation. In order to associate the employment and payroll records for all of the subsidiaries with the consolidated income tax return, a file had to be created containing the Corporation Affiliation Schedules (Forms 851), which list the Employer Identification Numbers of all of the subsidiaries for a given parent corporation.

The Form 851 file was linked on subsidiary EIN to the Forms 941/943 file, and the

appropriate payroll and employment amounts were appended to each subsidiary record which linked. The resulting file was then linked to the Corporation sample file using the parent EINs, and the amounts for each subsidiary linking to a Corporation record were added to payroll and employment fields appended to that record. The Corporation sample file (containing parent and nonconsolidated EINs) was then linked to the Forms 941/943 file, and payroll and employment, for each parent and nonconsolidated record which linked, were added to the payroll and employment fields appended to that record. (See Figure 3.)

FIGURE 3.--CORPORATION LINK PROCESSING



The accounting period ending date of the SOI Corporation record was then used to determine which quarterly Form 941 or annual Form 943 amounts were to be aggregated to arrive at the 1982 Fiscal Payroll and Employment amounts. For full-year returns, data from the preceding four quarters were taken; if the accounting period did not end evenly on a quarter, fractions were used to approximate the preceding four quarters. For part-year returns, an assumption was made that the return represented the six months prior to the accounting period ending date, and the previous two quarters of employment and payroll data were used. Some of the records on the file had accounting period ending dates such that the full period's data were not available on the Forms 941/943 file. These records were designated out-of-scope; 339 such records, representing 0.30 percent of Proxy Payroll (Salaries and Wages + Compensation of Officers) were so designated.

In order not to inflate the financial data for linked records with data from records which showed no Forms 941/943 employment activity for the period covered by the Form 1120 record, records with Fiscal Payroll and Employment equal to zero were designated as falsely linked. These records were grouped with the records which did not link. Together these two groups total 17,288 records, containing 2.70 percent of Proxy Payroll.

Finally, an attempt was made to identify falsely linked records. Records representing different reporting units may link due to an incorrect EIN on either the Form 941/943 record or the Form 1120 record. This error in the linking variable can be detected by comparing variables on each file having a known relationship. One such test, for the corporate study, is a comparison of Forms 941/943 Payroll with Proxy Payroll calculated from the Form 1120. In addition to simply comparing these two amounts, a comparison between Forms 941/943 Payroll and Form 1120 Total Deductions was made, on the assumption that some Payroll could be "hidden" in other deduction items. There were 780 linked records containing 0.20 percent of Proxy Payroll which had Fiscal Payroll amounts greater than both the Proxy Payroll and Total Deductions; these records were designated "falsely linked". The remaining 75,268 linked records, containing 96.8 percent of Proxy Payroll, were designated as "true links." These included any record representing a consolidated return for which the parent or any subsidiary linked to a Forms 941/943 record.

IMPUTATION AND REWEIGHTING

At this point, three of the five possible link outcomes (true link, true nonlink, and false link) have been addressed. The two remaining outcomes are a partial link, and a false nonlink proper [9]. (See Figure 2 for definitions.) Neither of these problems is as easy to deal with as a false link. Addressing these outcomes involves a two-stage process, identification and adjustment. In the case of the partially linked records, they may be directly identified by adopting an operational definition based on empirical research. This is not the case for the false nonlink proper, where an implied identification must be made and some adjustment to the linked records undertaken to account for false nonlinks.

Also, the adjustment procedures adopted for the two cases differ. It is useful to think of the partial link case as the analog of item

nonresponse in a survey, where some of the multiple Form 941 "blanks" are "filled in", while others are not. This suggests an item imputation strategy, while the false nonlink proper is analogous to a unit nonresponse, which suggests a reweighting approach [10].

Partial Match

This section will describe the treatment of partially matched Corporation records.

Identification of partial matches.--The first step in addressing the partial link problem was the identification of the partially linked records. By definition, the partially linked records were consolidated. Thus, the file was divided into consolidated and nonconsolidated subsets. The rest of the partial link definition, that some of the Forms 941 which represented members of the consolidated group had failed to link, was then employed. Given that the Fiscal Payroll (FP) reported on the Form 941/943 records and the Proxy Payroll (PX) reported on the Form 1120 are conceptually similar, it is reasonable to expect that the Fiscal Payroll/Proxy Payroll (FP/PX) ratio will be approximately one for a completely linked record and something less than one for a partially linked record.

A tabulation was prepared reflecting this reasoning, which showed the percentage of linked records within a given range of values of FP/PX from zero to two by increments of tenths. Significantly larger percentages of consolidated records than nonconsolidated records had values of FP/PX below the 0.7-0.8 range, while the percentages of records contained in the strata above this range were similar. Therefore, the partially linked records were operationally defined as those consolidated records with FP/PX less than 0.75.

Development of imputed amounts.--Next, an adjustment procedure was developed for these records. A ratio-based item imputation scheme was adopted. The first step was the definition of a donor set. Two primary problems exist in the Corporation linked file which cause Fiscal Payroll to be markedly different from Proxy Payroll. One is the partial link problem which causes the FP/PX ratio to be low. Another problem is the misreporting of payroll expenses in other deduction items. This would cause Proxy Payroll to be artificially low and, subsequently, the FP/PX ratio would be artificially high. In order not to choose any of these records as FP/PX ratio donors, donor records were limited to those with FP/PX between 0.75 and 1.50.

After the partially linked records and completely linked donors were identified, it was necessary to develop some method of associating a particular donor with a donee. The file was stratified into imputation cells according to the classes identified in Appendix II. Donors and donees were associated using the following metric function. First, each record to be imputed was associated with the donor records within the same imputation cell which had the same two-digit (SIC-based) industry code. From these records the donor which minimized the absolute value of the difference between the two records' proxy payroll values was chosen to supply an FP/PX ratio for use in developing an impute.

The actual imputed amounts were derived as follows:

Let: x_i be a partial observation
x_k be a complete observation
E_i be observed employment
FP_i be observed Fiscal Payroll
PX_i be observed Proxy Payroll
TD_i be observed Total Deductions
y_{pi}' be the value of the fiscal payroll imputed amounts
y_{ei}' be the value of the employment imputed amounts.

Let:

$$R_{pk} = \frac{FP_k}{PX_k}$$

$$R_{ek} = \frac{E_k}{PX_k}$$

$$\alpha_i = 1 - \frac{FP_i}{PX_i}$$

where R_{pk} and R_{ek} are payroll and employment imputation ratios from the donor record, α_i is a measure of the "missingness" of data in the ith partially linked record, and x_i and x_k are both contained in the Jth imputation cell. Then

$$y_{pi}' = \alpha_i R_{pk}(PX_i) + FP_i$$

and

$$y_{ei}' = \alpha_i R_{ek}(PX_i) + E_i$$

unless $\alpha_i R_{pk}(PX_i) + FP_i$ is greater than both PX_i and TD_i; then

$$y_{pi}' = PX_i$$

and

$$y_{ei}' = (\alpha_i R_{ek}(PX_i) + E_i) * (PX_i/(\alpha_i R_{pk}(PX_i) + FP_i)).$$

These imputed amounts were then appended to the partially linked records. Note that the imputation employed the false link test, that Fiscal Payroll must be less than or equal to Proxy Payroll or Total Deductions, as a bounding condition on the size of the imputed value [11].

False Nonlink Proper

The remaining link outcome, false nonlink proper, presented the most difficult identification problem. There were clearly categories of records, for example, those with large Proxy Payroll and Business Receipts, for which it might be reasonable to assume that all of the records should have linked. Indeed, this was a key assumption of our technique. However, this leaves a large gray area, namely, those records whose qualitative characteristics do not support such a powerful assumption. Does false

nonlink occur in these records as well? It seems likely that it does, but one is left with the question of how to identify those records which are falsely nonlinked. There is no simple answer to this question, therefore, a reweighting strategy was adopted which did not rely on being able to identify specific records.

Data reduction.--The initial stage of the adjustment process is to "identify" the falsely nonlinked records. The first step in pursuing this strategy was to create an analytical table predicting link status. For the 1982 Corporation Link Study, a 13 x 8 x 8 x 11 x 2 (Industry x Size of Total Assets x Size of Business Receipts x Size of Proxy Payroll x Link Status) table was created. While this table is too large to be practical for developing reweighting factors, it was used as a starting point for empirical analysis, using a contingency table approach, of the effects of each variable on link status. An APL computer routine, called CONTAB, was used to construct alternative tables to the analytical table under the assumption of simpler interactions between the predictive variables and link status. The alternative tables were compared to the original table using a relative distance measure based on the minimum discrimination information number (MDIN). First, a simplified model containing interactions between each of the predictor variables and link status was used to construct a table, and this table was compared to the original data table to generate a baseline MDIN. Next, four models, each omitting the interaction of one of the predictor variables and link status, were used to generate MDINs. These MDINs were then compared with the baseline MDIN. Any model which generated a significantly larger MDIN than the baseline model omitted important information. Conversely, if omitting the interaction of a predictive variable with link status changed the distribution of the data within the table very little, then that variable had little affect on link status. From the models, it was concluded that Proxy Payroll was the strongest indicator of link status and that Total Assets had the least association with link status. These results led to the collapsing of the 5-way table into a 4-way table exclusive of Assets.

After determining the least complex model which yielded an acceptable MDIN, we continued with the analysis of the classes within each variable which represented useful gradations of the variable. This was done according to two criteria. First, a routine known as EFFECTS was used. Using the table constructed by the CONTAB routine with the least complex model yielding an acceptable MDIN, EFFECTS employed logit analysis to produce a quantitative measure of the effect of each stratum of each variable on the distribution of data in the cells of the table. By using EFFECTS with link status and one other variable, it was possible to determine for which contiguous classes of the variable the effect of that variable on link status is similar. It was then possible to collapse these classes together, yielding a simplified table. The second criterion for collapsing is the presence or absence of a significant quantity of data in a region of the table. If a given class of a variable contains little or no data, this class may be collapsed without losing much information. If, on the other hand, two classes contain a great deal of data, it may be ill advised to collapse these two classes even given very similar effects.

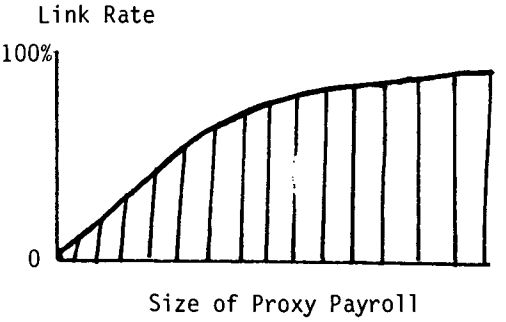
Identification of false nonlinks.--Following this analysis of each variable's effect on link status, and after collapsing the table, the next step was the development of the reweighting factors themselves. A key assumption was adopted at this point. For some regions of the table, in which Business Receipts 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.

Employing our assumption, we considered cells representing high Proxy Payroll classes to be excellent candidates for one hundred percent link status. (That is, a record for which an amount of proxy payroll greater than that which might indicate payments for casual labor is present likely represents a corporation which would be required to file an Employer's Quarterly Income Tax Return.) This idea is represented graphically in Figure 4.

The shaded area, with high Proxy Payroll and high link rate, represents the 100 percent area. Although the graph is only two-dimensional, another axis exists for each remaining variable, creating a four-dimensional surface, some region of which is a one hundred percent link region.

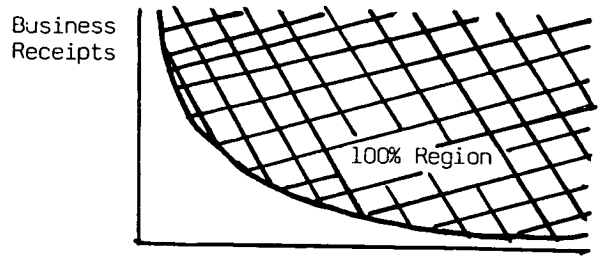
Figure 4.--An Example of Corporation Link Status By Size of Proxy Payroll

Amount of Proxy Payroll	Number of Records		Link Rate
	Linked	Nonlinked	
$0	1	49	.02
$1-10K	35	7	.83
$10-25K	81	9	.90
$25-50K	98	11	.90
$50-100K	115	10	.92
$100-500K	200	15	.93



Using this assumption, we proceeded industry-by-industry to develop the 100 percent link regions. Tables were produced classifying link rates by Proxy Payroll and Business Receipts. Examination of these tables revealed patterns similar to that in Figure 5 (with a slight variation for each industry).

Figure 5.--Determination of the 100 Percent Link Regions



This process identified one of the sets of records discussed earlier, 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.

Development of reweighting factors.--The development of reweighting factors may be conceptualized as follows. After the imputation procedure, it becomes appropriate to treat the partially linked and completely linked records as one category, simply designated "linked."

Assume 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. Let X_{Ai}, denote the sampled value of Ath item in the ith record, and w_i represent the weight determined by the rate at which returns in the record's class were sampled in the 1982 SOI Corporation study. Then

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

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

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

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

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

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

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

Effectively, the reweighting "subtracts" the false nonlinks from the other (true) nonlinks, and "adds" the false nonlinks to the links, where they belong.

Applying this method, 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 and nonlinked, 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 nonlink. Note that this was accomplished without the need for specific identification of the falsely nonlinked records. After this reweighting,the file was considered final; false links had been removed, partial links had been adjusted using imputation, and, finally, false nonlinks had been adjusted for by reweighting.

ANALYSIS OF THE 1982 DATA

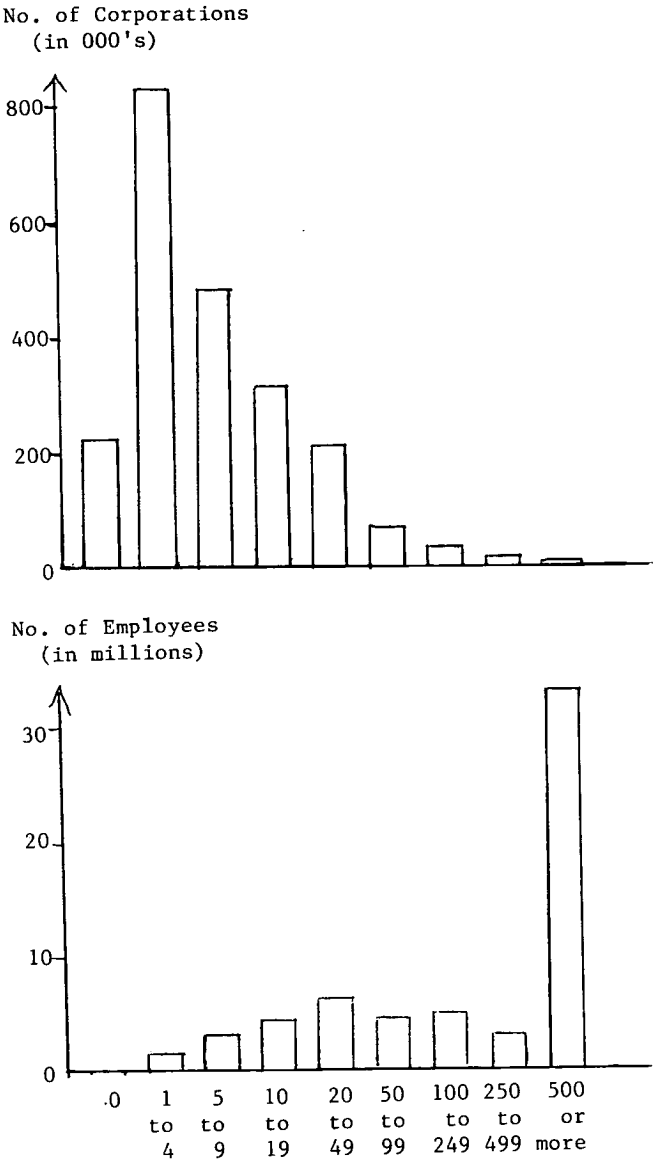
1982 Corporation Data

This analysis of corporation data will focus on examination of small versus large corporations which is the primary interest of the study's sponsor. For 1982 there were approximately 28 million people employed by over 2 million small corporations. (Small business here is defined as firms having under 500 employees.) Over 56 percent of the employment was in Wholesale and Retail Trade, and Services. The Manufacturing industry alone accounted for 22 percent of the employment in small corporations.

While 835 thousand or 38 percent of all corporations had 1-5 employees, with about 1.9 million people employed, the greatest number of employed were in corporations having 500 or more employees. Almost 33 million are employed by the 7,613 largest corporations. (See Figure 6.)

For all industries, average payroll was $15,400 for businesses with under 100 employees. The average payroll for companies with under 500 employees (small businesses) was $15,901. For all industries, average payroll

Figure 6.--Summary of 1982 Corporation Employment by Size of Employment



rises to $18,504 for businesses with 500 or more employees. The average payroll for businesses with under 5 employees was slightly higher with an average payroll for all industries of $19,904. Salaries would tend to be higher in this employment size class since employees would be more involved in management.

Across industries, payroll ranged from $6,268 for motion picture theaters and $6,740 in eating and drinking establishments, which are generally minimum or below minimum wage employers, to $35,452 for security and commodity brokers and $35,584 for offices of physicians. Ten percent of all corporations had no employment.

FUTURE ENHANCEMENTS

As noted previously, the methodological problems in all of these studies have been those of false links and, particularly, false nonlinks. Various algorithms have been devised to address the false link problem, primarily by comparing "similar" financial data elements and making a judgment as to whether they fall within an acceptable range. As greater familiarity with the data has been gained, these comparisons have been "fine-tuned" to some degree. If the comparison indicates a relationship that is thought to be "highly unlikely", the linked records are essentially treated as if a link did not occur.

The false nonlink problem remains the primary methodological concern, and various approaches have been attempted to deal with it. In the three 1979 studies, imputation procedures were developed, and each file was reweighted to attempt to account for this problem. However, while it was known which business returns did not link a payroll record, it was not clear how many of these were valid nonlinks because the business did not have payroll. For the imputation procedure developed, some assumptions were made on the likelihood that a link "should have occurred" and the files were reweighted accordingly.

In the 1982 studies, the Form 941 population or "universe" file was independently tabulated by type of business to derive control estimates of the number of businesses with payroll by type of business. These control estimates were then used in the reweighting process to ascertain the overall effects of various imputation adjustments. In future work, greater analysis of the Form 941 universe files is planned to improve their usefulness for developing control estimates and to examine the feasibility of deriving controls by industrial division.

In the 1982 partnership study, the EINs used to link the partnership and payroll files, were all taken from revenue processing records (i.e., returns that had been successfully posted to the Business Master File). Since "problem cases" in the processing of these records would have been addressed in revenue processing activities, the increase in "quality" of the linking variable was thought to possibly remove the need for false nonlink imputation in this study. However, significant nonlinks did occur anyway. At the present time, this situation is being investigated. In general, we believe that improvements in the quality of the linking variable could substantially reduce the need to commit such a significant level of resources to file reweighting. Furthermore, beginning with Tax Year 1985, SOI Corporation EINs will be generated from the revenue processing system.

While reduction of reliance on adjustment techniques is desirable, improvement of techniques used is also sought. Currently, multiple imputation techniques for adjustment of the partial links are being developed in order to reduce the additional variance introduced by imputation and to yield a measure of that variance.

Finally, in response to a need expressed by SBA, we are examining design issues in creating a panel of Form 941 entities to cover a multi-year period. Since this file would have indicators of type of business and industry, it could be analyzed to track the growth of

employment and payroll of individual firms. In addition, this file could be linked to the SOI business files on a periodic basis to enable a more detailed examination of the financial data. We believe that the creation of such a file would prove to be an invaluable resource in the growing field of business demography.

ACKNOWLEDGMENTS

The authors have greatly appreciated the encouragement and patience of Wendy Alvey, Beth Kilss, Kimm Bates, and Fritz Scheuren. Their suggestions and editing have helped to create a better paper. Both Wendy and Beth are to be congratulated for their colorful art work which brightened and simplified the presentation of the paper. The work and writings of Linda Taylor, Paul Rose, Nick Greenia, and Lock Oh are the foundations of this study and paper.

NOTES AND REFERENCES

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

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

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

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

[5] See Standard Industrial Classification Manual, Enterprise Standard Industrial Classification Manual, Office of Management and Budget, Statistical Policy Division, Washington, DC, 1972.

[6] For additional information, see also 1982 Corporation Income Tax Returns, Internal Revenue Service, 1985.

[7] For additional information, see also U.S. Internal Revenue Service, Statistics of Income Division, 1978-1982 Partnership Returns, Washington, DC, 1985.

[8] For additional information, see also U.S. Internal Revenue Service, Statistics of Income Division, 1982 Sole Proprietorship Returns, Washington, DC, 1983.

[9] Scheuren, Fritz and Oh, H. Lock, "Fiddling Around with Nonmatches and Mismatches," 1975 American Statistical Association Proceedings, Social Statistics Section, pp.627-633.

[10] "Reweight for missing records, impute for missing items." Per Little, Roderick J.A., Survey Nonresponse Adjustments, University of California, Los Angeles [Forthcoming].

[11] While cost constraints prevented its implementation for this study, a multiple imputation scheme, employing a metric function which maps each donee record to a neighborhood of the donor set, is a possible extension in future work.

BIBLIOGRAPHY

Greenia, Nick, "1979 Sole Proprietorship Employment and Payroll: Processing Methodology," Statistics of Income Division, IRS, 1982.

Hinkins, Susan M., "Imputation of Missing Items on Corporate Balance Sheets," 1982 American Statistical Association Proceedings, Section on Survey Research Methods, 1982, pp. 254-259.

Little, Roderick J. A., "Missing Data Adjustments in Large Surveys," Journal of Business and Economic Statistics, [Forthcoming].

Little, Roderick J. A. and Rubin, Donald B., Statistical Analysis with Missing Data, John Wiley & Sons, Inc. (New York), 1987.

Scheuren, Fritz, "Methodologic Issues in Linkage of Multiple Data Bases," Record Linkage Techniques--1985, Internal Revenue Service, 1985, pp. 155-178.

This appendix provides definitions on selected terms and concepts which are used throughout the paper.

Consolidated Returns.--Consolidated returns were income tax returns which contained the combined financial data of two or more corporations meeting the following requirements: (1) a common parent corporation owned at least 80 percent of the voting power of all classes of stock and at least 80 percent of each class of nonvoting stock (except stock which was limited and preferred as to dividends) of at least one member of the group; and (2) these same proportions of stock of each other member of the group were owned within the group.

Corporations electing to file consolidated returns in one year had to file consolidated returns in subsequent years, with certain exceptions. The consolidated filing privilege could be granted to all affiliated domestic corporations connected through stock ownership with a common parent corporation except: (1) regulated investment companies; (2) real estate investment trusts; (3) corporations for which an election to be treated as a possessions corporation under Code section 936(e) was in effect; (4) corporations designated tax-exempt under Code section 501; and (5) Domestic International Sales Corporations (DISC's). Under prior law, affiliated insurance companies were allowed to file a consolidated return if they were taxable under the same provisions of the Code. However, noninsurance companies with which they also may have been affiliated could not be included in the same return. Starting with taxable years beginning after December 31, 1980, insurance companies were allowed to file a consolidated return which included noninsurance companies as long as the noninsurance companies had been members of the affiliated group for 5 taxable years, that is, since January 1, 1976.

A consolidated return, filed by the common parent company, was treated as a unit, each statistical classification being determined on the basis of the combined data of the affiliated group. Therefore, filing changes to or from a consolidated return basis affect year-to-year comparability of certain statistics (such as data classified by industry and size of total assets).

Designation of Imputation Cells

After defining the partial link and donor groups, the next problem facing the analyst is the division of the linked file into groups of records which have similar payroll and employment characteristics. Four variables were used to do this; the following classes were used for stratification:

1. Proxy Payroll size classes:

   $1 under 10,000
   $10,000 under 100,000
   $100,000 and over

2. Business Receipts size classes:

   $0 under 100,000
   $100,000 and over

3. Total Assets size classes:

   Less than $5,000,000
   $5,000,000 and over

4. Industries

   Agriculture, Forestry, and Fishing
   Mining
   Construction
   Manufacturing
   Transportation, Communication, and Public Utilities
   Wholesale Trade
   Retail Trade and Wholesale and Retail Trade Not Allocable
   Finance
   Insurance
   Real Estate
   Holding Companies
   Services
   Nature of Business Not Allocable