RECORD LINKAGE AND IMPUTATION STRATEGIES IN THE 1982 CORPORATION EMPLOYMENT AND PAYROLL STUDY

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 sample of corporation returns is a rich source of data on business activity, containing detailed income statement and balance sheet data. This file, however, contains only two of the three frequently-used measures of the size of a business, receipts and assets. It does not contain the third measure, employment, because this is not reported on corporation 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 business 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 affect 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 [1,2,3,4,9].

The current paper will look at a number of different aspects of the Tax Year 1982 IRS Corporation employment and payroll link study. Organizationally, the paper is divided into four parts. The first part describes the sources of data used in the study. The second part provides a detailed discussion of the Corporation linking methodology. This is followed by a description of imputation and reweighting for partial links and false nonlinks. 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 file used in this study is the IRS' Statistics of Income (SOI) sample of Corporations for Tax Year 1982. The data for employment and payroll added to the file 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.

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 is 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.

Employment Tax Returns

Employment tax returns, Forms 941 series and Forms 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 and an aggregate payroll figure for the period of the return.

CORPORATION LINK PROCESSING

Defining a Linkage

The linking of employment returns to corporation returns is carried out on a record-by-record basis using the Employer Identification Number (EIN) as the linking variable. A typology of linking outcomes is provided in Figure 1. Probably the most critical element in any record linkage is defining a true link. 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

Figure 1.--Possible Link Outcomes

TRUE LINK.--A link between a Corporation record and an Employment tax return record representing the same reporting unit.

TRUE NONLINK.--A Corporation 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.--A Corporation record which fails to link, and for which an Employment tax return record representing the same reporting unit exists.

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.

was required that the records represented different reporting units before they were designated as falsely linked. In fact, this was only done when the linked records 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, and sometimes noncalendar, periods, while the non-agricultural employment returns were filed for calendar quarters. This difference in reporting periods 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, Forms 941/943 records with no nonzero data for the accounting period of the corresponding corporation return were excluded from the range of our linking function. Unfortunately, these records were grouped with the true nonlinks, and no independent count of their number or total of their Proxy Payroll (the sum of Salaries and Wages and Compensation of Officers as reported on Form 1120) is available. The linking domain was also restricted. The sample file contained records which represented prior-years' returns acting as proxies for current year 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 excluded from the true link category. These records were designated out-of-scope; 339 records, representing 0.30 percent of Proxy Payroll were so designated.

Linking the Returns

The Corporation Employment Link study is made more complex by the presence of consolidated returns. Consolidated returns are filed by corporations which own another corporation [5]. 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 lists the Employer Identification Numbers of all of the subsidiaries for a given parent corporation. (See Figure 2.)

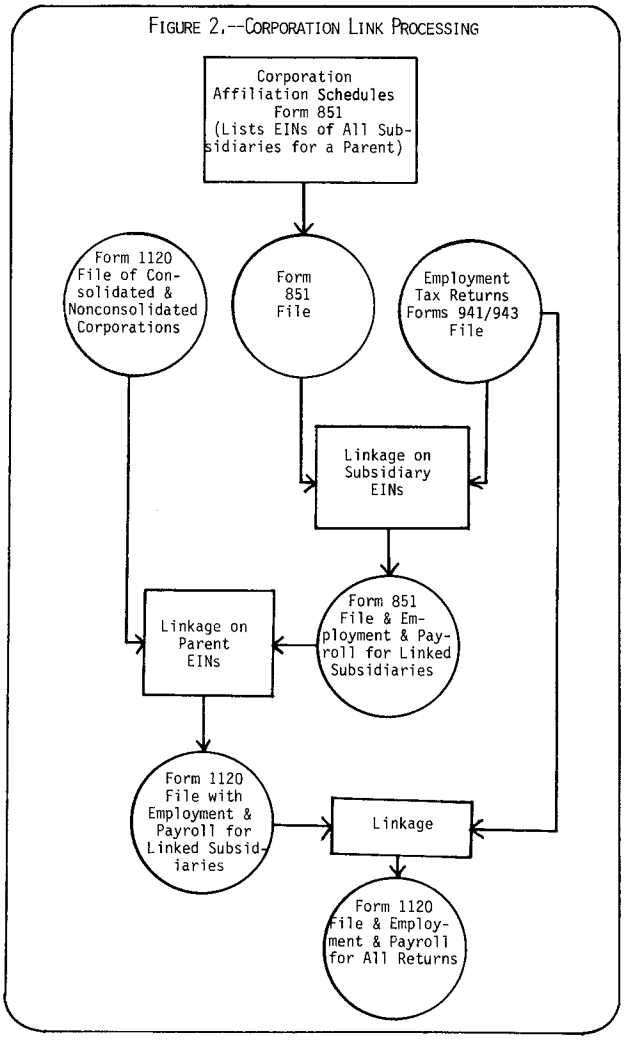

FIGURE 2.--CORPORATION LINK PROCESSING

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.

Results of 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 appropriate 1982 Fiscal Payroll and Employment amounts that are consistent with the financial reporting of the corporation. 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.

Next, 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 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 0.2 percent of Proxy Payroll which had Fiscal Payroll amounts greater than both the Proxy Payroll and Total Deductions; these records were designated "falsely linked". (See Figure 3.)

Figure 3.--Results of Link Processing

Link Outcomes	Number of Returns	Percent of Proxy Payroll
Total Sample.....	93,675	100.0
True Links.....	75,268	96.8
True Nonlinks*...	17,288	2.7
False Links.....	780	0.2
Out-of-Scope.....	339	0.3

*This category includes linked returns for which there was no nonzero Form 941/943 data corresponding to the Form 1120 accounting period.

The remaining 75,268 linked records, containing 96.8 percent of Proxy Payroll, were desnated 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. Finally, the true nonlinks (including the linked records with no nonzero data for employment and payroll) totaled 17,288 records, containing 2.7 percent of the Proxy Payroll.

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. (See Figure 1 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 [6].

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, reported on the Form 941/943 records, and the Proxy Payroll, 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 the partially linked 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 industrial division and the size of several key variables. 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 industry code [7]. 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 imputed amount.

The actual imputed amounts were computed as follows:

Let: x_i be a partial observation
x_k be a complete observation
E be observed employment
FP be observed Fiscal Payroll
PX be observed Proxy Payroll
TD be observed Total Deductions
y_{pi}' be the value of the fiscal payroll imputed amount
y_{ei}' be the value of the employment imputed amount

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

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

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

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. The imputed amounts were constructed as follows:

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

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 [8].

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 to the Form 941/943 file. 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 nonlinking 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.

Next, 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 effect 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, the study 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 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.

In light of this assumption, cells representing high Proxy Payroll classes were considered to be excellent candidates for 100 percent link status. (That is, a corporation reporting high payroll expenses on Form 1120 is likely to file a Form 941.) Tables were produced classifying link rates by Proxy Payroll and Business Receipts. Examination of these tables revealed a pattern of cells with match rates exceeding 90 percent in the high Proxy Payroll-high Business Receipts region. These areas were designated 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 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. (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.

FUTURE ENHANCEMENTS AND RELATED RESEARCH

As noted previously, the methodological problems in this study 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. Research is underway on the effect of the imputation and reweighting procedures on the distribution of the data by industry and size of employment. The results of this research are expected to lead to further refinements of our adjustment techniques. A proposal which is already under consideration is the adoption of a multiple imputation technique which will allow us to estimate variance due to imputation.

Taking an opposite tack, an earlier proposal, which has since been implemented, required subsidiary corporations to report the EINs of their parent corporations on their Employment tax returns. It is hoped that this will provide us with a more reliable and less costly method of associating subsidiaries and parents than the current use of the Form 851 file.

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, 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, "Tax Year 1979 Partnership Employment and Payroll." 1978-82 Partnership Returns, Internal Revenue Service, 1985, pp. 221-236.

[5] For additional information, see also 1982 Corporation Income Tax Returns, Internal Revenue Service, 1983.

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

[7] Industry codes were based on the 1972 revision of the Standard Industrial Classification Manual.

[8] 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.

[9] Moglen, Gail; Day, Charles; and Petska, Tom, "Record Linkage and Imputation Strategies in the 1982 Business Employment and Payroll Studies." Statistics of Income and Related Administrative Record Research: 1986-1987, Internal Revenue Service, 1987.

BIBLIOGRAPHY

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, pp. 155-178, 1985.