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

Douglas K. Sater, Bureau of the Census

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

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

I. INTRODUCTION

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

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

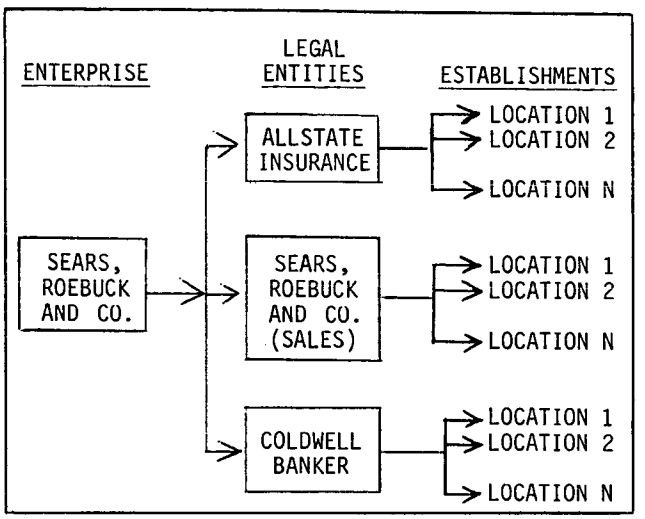
The addition of economic data to the SIPP will enable researchers to obtain improved estimates of the impact of economic and institutional forces which have been intensively studied but are only partially understood or measured. Some of the areas in which the matched file can yield new insights are: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

II. DEFINITIONS

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

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

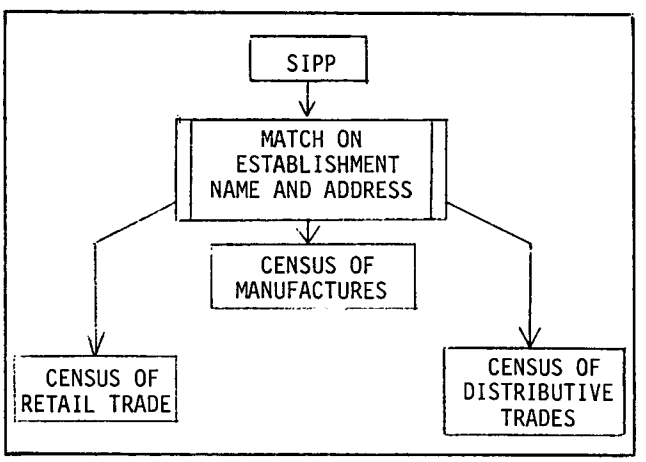
Figure 1.--A Partial Example of Basic Definitions



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

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

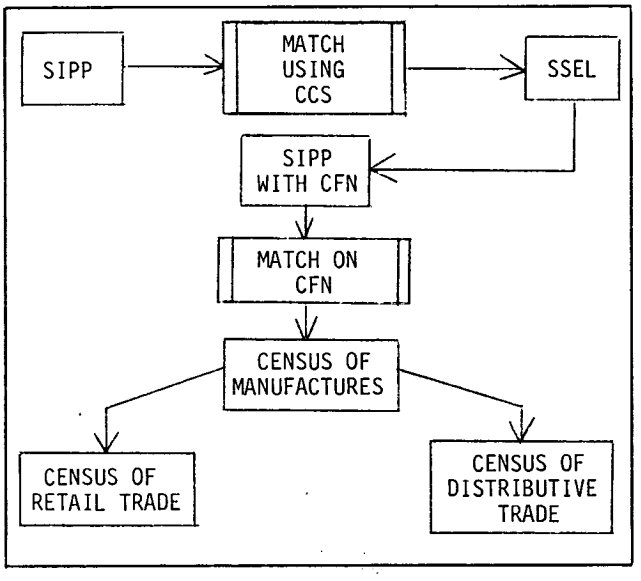
Figure 2.--Simplified Diagram of Direct Match to Three Economic Censuses



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

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

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



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

The CCS operates in two basic modes:

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

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

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

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

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

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

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

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

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

III. MATCHING PROBLEMS

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

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

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

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

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

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

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

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

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

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

IV. PRE-TEST RESULTS

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

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

Table 1.--Results of Address Search Operation

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

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

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

Table 2.--Results of Matching Test

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

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

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

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

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

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

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

V. OTHER ISSUES

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

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

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

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

4. Errors in EIN's.

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

6. Suppression issues in data releases.

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

BIBLIOGRAPHY

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

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

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

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

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

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