IMPROVING INDUSTRY AND PLACE-OF-WORK CODING IN THE CONTINUOUS WORK HISTORY SAMPLES

Bruce Levine, Bureau of Economic Analysis

During recent years, there has been an increasing level of use of the Social Security Administration's (SSA) Continuous Work History Samples (CWHS). A partial bibliography compiled in 1976 listed 68 separate studies using the CWHS.1/ Since that time, even though legal complications have prohibited the release of files of years subsequent to the 1975 data year, the frequency of application of this source of data has increased.

The original purpose of the CWHS was to gather statistics from the material already being collected for internal SSA monitoring. The data is generated from records that are necessary for the administration of SSA programs. The expense of assembling the CWHS is, therefore, mostly the marginal costs of processing the inclusion of variables not essential to SSA such as employee's place-of-work and industry.

A forthcoming report by the Subcommittee on Statistical Uses of Administrative Records states, "Statistical use of administrative records grew rapidly during the 1970's, in large part as a response to legislative requirements for timely data to use in the distribution of Federal funds to State and local governments. The principal reason for increasing reliance on administrative records for statistical data is the availability of administrative records which can be used to obtain small area data at a minimal cost and without increasing respondent burden. And cost is likely to be an increasingly important factor in the statistical use of administrative records in the 1980's."2/

The kinds of uses which have been made of the CWHS are many and varied. Demographic and industrial information enable cross-sectional analysis by area and the ability to study the structure of the workforce. The longitudinal nature of the file allows analyses of workforce change including geographic and industrial mobility. The recent incorporation of place-of-residence codes has brought about the ability to study commuting flows and population migration.

Among potential future inclusions that might be added to enhance the usefulness of the CWHS are variables from Internal Revenue Service records of taxation, National Center for Health Statistics, SSA's Supplemental Security Income program and the Medicare program. These will improve the ability to do epidemiological research and to study mortality and morbidity. Also other demographic variables such as occupation and marital status, may be incorporated in the file.

The validity of the analytical results based on the CWHS has recently been questioned. As is the case with any data source, the CWHS is not perfect. The limitations of incomplete workforce coverage and sampling variability have been documented and are well recognized by those using the files. The results of a major study by David Cartwright using the 1975 10-percent CWHS have focused much attention on employer reporting errors which lead to incorrect place-of-work and industry information. This problem was considerably more widespread than had previously been thought. He estimated that over 11 percent of all workers were probably miscoded by place-of-work.3/

In a recent memorandum, Henry Patt, Director of SSA's Division of Statistics states, "The quality of industry and geographic data in the CWHS has been deteriorating over the years, primarily due to depleted staff in key areas dealing with obtaining this information from employers, classifying it properly and maintaining it in our files. I am concerned that this deterioration will reach a point (if not already there) that the CWHS will be unusable for certain purposes."4/

In as much as place-of-work and industry is more important to the work of other researchers than to internal SSA administrative programs, interagency cooperation has been necessary. The purpose of this paper is to report the progress of an ongoing effort by SSA and the Bureau of Economic Analysis (BEA) to evaluate, correct, and to improve future procedures to include geographic and industrial information in the CWHS.

Timing of File Creation

The annual Employee-Employer (EE-ER) files are created in their final form 1½ - 2½ years after the end of the data year. To meet a demand to release data on a more timely basis, SSA has supplied the Census Bureau and BEA with preliminary first quarter files which are extracted approximately 1 year after the end of the quarter. The preliminary files are drawn as soon as all necessary information is accumulated. SSA has no opportunity to verify or to resolve problems in any of the included data. Consequently, there is a much higher proportion of geographic and industry codes that cannot be classified or that are incorrectly coded in the preliminary files.

To determine the effects of timing of the extraction of the files, the 1975 preliminary first quarter file was compared to the final first quarter data drawn from the annual 1975 EE-ER.

Workers present in the final file but not in the preliminary are individuals for whom employers filed late quarterly wage reports or for whom SSA delayed processing because of ambiguous information on the report that required follow-up contact. Workers in this category have considerably below average wages. This probably means that many of them are part-time or intermittent workers. There is a disproportionate number of females and the average age is lower than among other groups. Many of these individuals may be employed by firms that report late to SSA for financial reasons.

While the effort to isolate the characteristics of workers unclassified in the preliminary and classified in the final file and those from whom information change between the two files is still underway, some hypotheses can be stated.

Individuals who were not classified in the preliminary file but who were in the final file are probably mostly workers whose employer reported them under a new employer identification number or reporting unit number. These were added to the SSA files after the preliminary, but

before the annual file was created. New reporting unit designations can occur because of business births or because of reorganizations of existing companies. Tabulations for the 1975 comparison show wages higher than average for this category of worker. This suggests that delayed information on reorganizations of existing large, typically high wage companies has a greater effect than that of the late addition of business births which are usually companies who pay below average wages.

The fact that some workers had valid geographic and/or industrial codes in the preliminary file but had different codes in the final file is probably because updated information for a given employer identification number or reporting unit was included after the creation of the preliminary file. The two most important sources of updated information are:

1. Reports that SSA requests from employers upon learning of a change in company structure.
2. New codes obtained from the Census Bureau in a periodic match of SSA records to files from the economic censuses.

The reports that SSA receives to update their files are probably mostly for larger companies. New information included from the censuses, however, is mostly on smaller and single-unit employers. The workers who had different State codes in the preliminary and the final 1975 files have above average wages, but, generally they are lower than workers wages whose State became classified in the final. This wage pattern combined with the fact that about 3 percent of the workers changed State from one file to the other indicates that an update with Census information was done between the extraction of the preliminary file and the final 1975 files.

Presumably the changes in place-of-work and industry occurring between the creation of the two files would result in more accurate information. In some cases, however, this is not true. For example, an employer could file information reporting a change in the location of his business that occurred in the third quarter of 1975. If the change were received before SSA updates its files in late 1976, it would be included, making the first quarter 1975 place-of-work incorrect.

The Employer File

A key element in the evaluation and eventual improvement of the EE-ER files is the study of the reporting patterns of employers. Firms that operate several establishments are requested to comply with the Establishment Reporting Plan (ERP). This voluntary plan asks that employers report workers grouped by establishment. Establishments in the same county with the same kind of industrial activity are requested to be reported as if they were one unit.

Those working with any of the CWHS files have been aware that the lack of adequate compliance with the ERP has caused many problems in successfully using them. In his 1978 study, David Cartwright noted the following kinds of difficulties with the reporting of multi-unit employers.

1. Some refuse to comply with the voluntary ERP.
2. Establishments may be renumbered by employers without SSA knowledge.
3. Employers report the workers of more than one establishment under one establishment number.5/

Because of sample size limitations, there has really been no effort in the past to evaluate the reporting patterns of single-unit and small multi-unit employers.

An attempt is currently underway to analyze the reporting of all employers. The 1973, 74, and 75 EE-ER files are being used in connection with this effort. Not only are these the most recent files to which access has been given, but they also include place-of-residence information from IRS records. The comparison of residence with work geography of individual workers serves as the major evaluation vehicle for place-of-work coding.

Aggregate tabulations of data by reporting unit characteristics will be designed to answer such questions as the extent to which commuting ratios are different among small, medium, and large employers; single and multi-unit employers; or are different by geographic region or by industry.

Table 1 shows some intermediate results based on the 1975 EE-ER file. It is quite apparent that there is more commuting and much more long distance commuting (residing in one State and working in another) among the employees of large firms. Some of the difference may be bonafide because larger employers pay higher wages and, thus, their workers can afford to commute. The higher ratios among larger employers also certainly indicate that the place-of-work reporting is not as accurate as that of smaller employers.

The contrast between single and multi-unit employers is not as striking. The overall commuting ratios are very similar between the two groups. More differences may be shown when the employer file mentioned above is assembled and a distance of commuting is substituted for the same vs. other State criteria.

Table 2 indicates SSA's ability to classify place-of-work by employer characteristics. The larger an employer is, the greater the probability that SSA will not be able to determine his place-of-work. There is also a very striking difference between single and multi-unit employers. The previously mentioned difficulties with the ERP cause an inability to classify multi-unit firms.

Curiously, there was a considerably lower proportion of employees of single-unit companies that could be matched with a place-of-residence. This is probably because these firms hire many intermittent and part-time workers. Over 50 percent of these employees had no wages during the first quarter and made less than $1,000 during the entire year. Also, over 50 percent of this group was less than 25 years of age, and, thus, may have had no history of paying income tax.

Aggregate tabulations of wages for various sizes of reporting units by industry and geographic region will be compared with County Business Pattern (CBP) data by establishment size to determine if problems such as poor compliance with the ERP by large employers lead to distortions of CWHS employment and payroll data.

Table 1

Percent of Workers Who Commuted Within a State and to Other States: by Size and Type of Employer

Employer Size	All Employers			Single-Unit Employers			Multi-Unit Employers		
	Number of Jobs	% Commuters Same State	% Commuters Other State	Number of Jobs	% Commuters Same State	% Commuters Other State	Number of Jobs	% Commuters Same State	% Commuters Other State
All Employer	1,028,043	25.0	12.5	742,510	23.6	13.2	285,533	28.6	10.5
1 Job	324,548	21.3	8.2	295,727	20.7	7.8	28,821	28.3	12.5
2 Jobs	142,455	23.8	10.6	117,936	23.3	10.5	24,519	26.0	11.0
3 Jobs	80,958	24.5	12.2	61,347	24.5	12.5	19,611	24.7	11.1
4 Jobs	50,202	28.0	14.0	39,963	25.3	13.2	16,239	24.4	10.5
5 Jobs	40,977	26.2	12.7	26,951	26.5	14.2	14,026	25.6	9.9
6 Jobs	31,865	25.7	13.6	20,262	26.3	15.8	11,603	24.6	9.8
7 Jobs	25,664	25.3	14.0	13,147	26.2	17.0	8,891	24.0	9.6
8 Jobs	22,038	25.3	14.0	13,147	26.2	17.0	8,891	24.0	9.6
9 Jobs	17,541	26.4	14.2	10,419	26.8	17.1	7,122	25.7	9.8
10-19 Jobs	97,837	26.4	14.4	53,739	26.2	17.9	44,098	26.6	10.1
20-49 Jobs	84,549	27.8	16.7	42,161	26.4	23.2	42,388	29.2	10.4
50-99 Jobs	40,155	27.3	18.6	17,803	25.5	28.1	22,352	28.8	10.9
100+ Jobs	63,254	37.5	23.9	27,314	32.6	43.1	35,940	41.3	9.3

Table 2

Percent of Unclassified Workers by Size and Type of Employer

	All Employers		Single-Unit Employers		Multi-Unit Employers	
	% Unclassified		% Unclassified		% Unclassified	
	place-work	place-res	place-work	place-res	place-work	place-res
All Employers	5.4	16.8	1.7	19.1	14.5	11.2
1 Job	5.9	18.1	2.6	19.1	31.0	10.5
2 Jobs	4.6	17.4	1.1	18.9	19.2	10.9
3 Jobs	4.6	16.9	.9	19.0	15.1	10.8
4 Jobs	4.1	16.7	.8	19.2	12.1	10.7
5 Jobs	4.6	16.3	.7	19.3	11.9	10.8
6 Jobs	4.7	16.8	.8	19.8	11.3	11.8
7 Jobs	5.0	15.6	.9	18.8	11.1	10.7
8 Jobs	4.6	15.4	1.2	18.9	9.7	10.2
9 Jobs	7.5	15.2	.8	19.1	16.5	9.9
10-19 Jobs	5.0	15.2	1.1	19.0	9.7	10.5
20-49 Jobs	5.8	14.8	1.5	18.7	10.0	11.0
50-99 Jobs	6.6	14.7	2.6	19.4	9.8	10.9
100+ Jobs	7.3	16.1	.8	19.1	11.9	14.1

Source: 1975 EE-ER File

Linking Employer Files Longitudinally

Once the employer files of each of the three years have been prepared they will be linked longitudinally. This will enable the examination of such phenomena as reporting unit births and deaths, major changes in reporting unit size, changes in geographic or industry coding for specific reporting units, and significant changes in worker commuting patterns in reporting units. The extent of commuting and long distance commuting will be used to help identify particular kinds of reporting problems and tendencies for such problems to appear, to remain stable, or to be corrected over time. In addition, where large reporting units are born or die, the workers from those units can be traced longitudinally to try to identify such phenomena as reporting unit renumbering, consolidation or deconsolidation of reporting units, and changes in employer identification numbers.

A test of the linked employer files has been accomplished examining establishments located in Michigan during either 1973, 1974 or 1975. Many small reporting units were represented in one year's sample but not in others. There were about 23,000 units reporting in the 1974 one-percent EE-ER. Of those, over 10,000 reported no employees in the 1973 sample and almost 8,000 had none in 1975. These establishments, on average were less than 50 percent of the size of establishments that had employees represented in more than one of the study years.

While a sizeable number of small establishments would be expected to be omitted from one year to the next, further examination of some of these cases is necessary. Multi-unit employers identifying their reporting units differently from year to year will cause this phenomenon. A routine method for identifying such employers will be used and with the aid of SSA, establishment identification will be studied.

A second problem category was establishments coded in different places from file to file. The Michigan test showed 254 establishments in 1974 that had been recorded in another State in 1973 and 89 reporting units in a different State in 1975. There were about 1,000 establishments that were coded in different counties within Michigan in 1973 and 1974, and 300 in 1974 and 1975. It is not likely that many of these reporting units actually moved. A file of all establishments of multi-unit employers that were reported with different geography as well as the larger single unit employers in this category will be jointly studied by BEA and SSA to resolve as many differences as possible.

There were 70 units that were not classified by area in 1973 that were coded in Michigan in 1974 and 9 that were in Michigan in 1974 and not classified in 1975. A routine correction can be made by assigning the area, classified areas that these units were reported into for the years that they were unclassified.

It is intended that the files prepared from the 1975 10-percent CWHS used in David Cartwright's 1978 study will be linked in to permit a more detailed examination of small employers. Also, inasmuch as the 1975 10-percent files were preliminary, more information can be gained about the effects of the timing of file extraction.

The CWHS Self-Employed (SE) File

SSA assembles an annual SE file in an analogous manner to the preparation of the annual EE-ER files. The SE file includes a record for everyone who submits a self-employment (SE) tax schedule with his or her Form 1040. The data items for each SE worker include race, sex, SE income, industry, and place of work. The sampling methodology is the same as that used with the EE-ER file. That is, social security numbers ending with the same digits are selected for both samples.

With some exceptions 6/, the SE files have not been used extensively for research purposes outside of SSA in the past. There are, however, numerous potential applications to the research community. The Small Business Administration, for example, has recently requested that a longitudinal sample file of individuals who had been self-employed during some part of the 1960-75 period be assembled. This file also includes longitudinal wage and salary information for SE workers drawn from the EE-ER files. This data will facilitate the study of individual proprietors and partners whose businesses are typically small. Other sources of such data are virtually nonexistent.

The total number of self-employed workers has a trend that is approximately countercyclical to that of the general economy. During bad economic times, workers in some industries lose their jobs or have their hours cut back; and thus, they rely on their own resources to live. As conditions improve, they return to the security of wage and salary employment.

There are relatively few blacks engaged in self-employment. Their concentration in declining center city areas may explain this lack of growth as compared to white workers. The political climate of encouraging the recruitment of blacks in large organizations and government may have also restricted the potential number of black self-employed.

There was a significant growth in the number of self-employed white females. Their growth in average income, however, has been less than the growth for all workers. Since many of these females are new to self-employment, their lack of experience could be the cause of lower earnings. Also, women may be entering part-time self-employment and low income fields such as child care.

Study of the SE files is enabling the formulation and testing of procedures to evaluate and to improve the main EE-ER files. The characteristics and inadequacies of the two data sets are, to some extent, similar and the number of cases in each year's SE sample is much smaller.

The Unclassified Industry Problem

Because the number of workers in the SE sample is so small (ranging from 60 to 68 thousand workers annually), it is most desirable to include as many as possible in any analysis. From 1960 through 1967, the level of industrially unclassified workers was manageable. There were 5 percent or less unclassified in each year. This rate was very good considering that no attempt is made to industrially code the partnerships included. Starting in 1968, however, there were far more individuals who could not be assigned an industry. The reason for this slippage in coding is that before 1968

SSA received the schedule SE's from IRS and assembled the file as a routine part of CWHS processing. Subsequent to 1968, however, IRS began to transmit the SE data on magnetic tape and problem resolution was difficult or impossible.

When the increasing magnitude of this problem was determined, it was clear that for the SE files to be of maximum utility a method of imputing classified industry codes for previously unclassified codes should be undertaken. A three phase approach has been formulated.

Step 1: SE Longitudinal Imputation

The most relevant information that can be used to determine what industry code to attribute to an unclassified is industry codes of other years of self-employed activity.

I_t = Industry code in year t which is unclassified.

If $I_{t-1} = I_{t+1}$ and both codes are classified, it would seem a safe procedure to substitute I_{t-1} for I_t. There were over 4,000 such imputations made in the 1960-75 SE longitudinal file.

If I_{t-1} is classified and I_{t+1} is either unclassified or not active, substitute I_{t-1} for I_t.

If I_{t+1} is classified and I_{t-1} is either unclassified or not active, substitute I_{t+1} for I_t.

If $I_{t-1} \neq I_{t+1}$ and both codes are classified, the procedure is not so clear cut. This situation is considered a tie, and, therefore, tie breaking criteria must be introduced.

S_t = The State code in year t. It must be classified to qualify for tie breaking. It is reasoned that if a person were in the same State during a year he was unclassified industrially as he was in an adjacent year when he was classified, his industry probably did not change. Therefore:

If $S_t = S_{t-1}$ and $S_t \neq S_{t+1}$, I_{t-1} would be substituted for I_t.

If $S_t = S_{t+1}$ and $S_t \neq S_{t-1}$, I_{t+1} would be substituted for I_t.

This tie breaking criteria cannot be employed if:
S_t is unclassified
or S_{t+1} and S_{t-1} are both unclassified
or $S_{t+1} = S_{t-1}$

For cases still tied the search is expanded to industry codes 2 years away.

If $I_{t+2} = I_{t+1}$ and $I_{t-2} \neq I_{t-1}$ substitute I_{t+1} for I_t.

If $I_{t-2} = I_{t-1}$ and $I_{t+2} \neq I_{t+1}$ substitute I_{t-1} for I_t

If the tie is still not resolved, the search is widened until all of the longitudinal information is exhausted. This exhaustive method may cause some imputations of industry codes that are suspect. For instance, a worker who was coded retail in 1962 and 1963 was not active in the SE file again until 1973 when his industry code was not classified. This procedure would impute a retail industry code for him in 1973. While a ten-year absence from self-employment could well indicate changes in working patterns, it is still likely that this worker is in the same or a related field.

The SE longitudinal imputation procedure has been implemented and the results are encouraging. In all years from 1961 to 1972, over 75 percent of the unclassified industry codes could be imputed. The end years, 1960 and 1973-1975, could not be imputed as successfully because there is less or no chance that longitudinal information will exist on both sides of the unclassified industry code. In addition, the 1973, 74 and 75 files had, by far, the most unclassified workers before correction.

The industrial distribution of the imputations made is also promising. The industrial patterns of imputations is reasonably similar to the distributions of industries as originally classified. The major exception, SIC 01 (farmers), is explainable. Since a different portion of the Schedule SE is specified for farm earnings than for all other, it is easier to determine whether an individual is a farmer. Consequently, it is reasonable that a lesser proportion of unclassified workers are actually farmers than is the percentage of classified workers.

Step 2: Wage and Salary Imputation

After the SE longitudinal imputation was performed, there were still a large number of industry codes that remained unclassified. There were over 17,000 individuals who were active in one or more years and never were classified by industry, thus making the longitudinal imputation impossible. This included four workers who were active in all 16 years and not industrially classified in any of them.

At this point longitudinal wage and salary information was brought into the procedure. The underlying theory was that in some industrial classifications, it would be likely that a person be employed in the same industry in which he is self-employed. There are only a relatively small number of industries can be classified in this manner. Among those that cannot be classified with the use of this method are:

1. SE workers who did not hold a wage and salary job during the year they were unclassified. This amounted to over 40 percent of the SE workers in 1970.
2. Unclassified SE workers who were also unclassified in the EE-ER file during the same year.
3. Those who were in an EE-ER industry that probably would not be the same as their SE industry such as automobile manufacturing or government.

A pilot study was done to determine the feasibility of transferring selected industry codes for an individual from the EE-ER where his code was unclassified for the same year in SE. The aim of the test was to relate individuals who were active in both files for a given year and compare their industry codes. The results were discouraging. There were, in every industry tested, 5-10 times more workers who had different EE-ER and SE industry codes than whose codes were the same. This makes such a correction highly questionable for most industries. More exhaustive testing for some industries such as construction workers, doctors and lawyers is still in process.

Step 3: The Earnings and Demographic Imputation

All of the industry codes that remain unclassified after steps 1 and 2 will be classified using this method. Sets of industrial probabilities will be calculated with regard to worker earnings, race, sex and age. A stochastic method will then be employed to distribute industry codes according to the probability distributions.

The Unclassified Place-of-Work Problem

In addition to the unclassified industry problem in the SE file, there is a growing number of workers with unclassified place-of-work. While the magnitude of the problem is not as large as that of unclassified industries, it is serious enough to warrant corrective action.

A three step strategy to classify place-of-work almost identical to the one for industry, will be employed.

1. If longitudinal SE geographic information is present, it will be used for imputation of place-of-work codes. If there is equal reason to code the worker to two places, similar tie breaking methodology will be introduced.
2. For places-of-work still unclassified after step 1, the appropriate State and county from the EE-ER file will be used. This will be more effective than the comparable step was for classifying industry because it can be assumed that if a person is employed in an area he will be self-employed in the same area at the same time.
3. A method to stochastically assign place-of-work codes based on income and demographic probabilities will be used to assign the remaining unclassified place-of-work codes.

Concluding Comments

While the effort to evaluate and improve industrial and geographic coding in the CWHS is far from finished, some significant progress has been made. The 1973-75 Employer File is almost complete. This will be a very powerful tool for the assessment of employer reporting. In addition, the correction procedures developed using the SE data provide a positive beginning for the improvement of file coding.

It will be vital to be cognizant of and to be able to deal with errors and inconsistencies in historical CWHS files because a new set of problems are being introduced. Starting with the 1978 data year a new method of reporting to SSA has been required of employers. SSA anticipates that this change will further erode employer participation in the ERP.

The current and potential increased value of this data to researchers and policymakers make it necessary that improvement efforts be continued. The growing demand for the demographic and longitudinal statistics for subnational areas is likely to make this a much more important data source in the future.

Footnotes

The views expressed in this paper are the responsibility of the author and not necessarily those of the Bureau of Economic Analysis or the Department of Commerce.

1/ Regional Workforce Characteristics and Migration Data, pp. 119-128, U.S. Department of Commerce, December 1976.

2/ The Subcommittee on Statistical Uses of Administrative Records of the Federal Committee on Statistical Methodology, Report on Statistical Uses of Administrative Records, p. 1, forthcoming.

3/ David W. Cartwright, Major Limitations of CWHS Files and Prospects for Improvement, Table 1, Presented at NBER Workshop on Policy Analysis with Social Security Research Files, March 17, 1978.

4/ Henry Patt, "CWHS Industry and Geographic Data," From a memorandum to John J. Carroll, November 1, 1979.

5/ Cartwright, op. cit, pp. 3-4.

6/ Dale E. Hathaway and Brian B. Perkins, "Farm Labor Mobility, Migration, and Income Distribution". American Journal of Agricultural Economics, pp. 342-356, May, 1968.