Occupation and Industry Data from Tax Year 1993 Individual Tax Returns

Peter Sailer and Terry Nuriddin, Internal Revenue Service
P.O. Box 2608, Washington, DC 20013-2608

Key Words: Administrative records, occupation, industry

For Tax Year 1993, the Statistics of Income Division (SOI) created a more elaborate database of individual income tax data than ever before. It contained not only a sample of individual income tax returns, but also matching information documents of every description—documents filed by the taxpayers' employers, banks, brokerage houses, pension funds, etc. In addition, through matches to other administrative files, we gender-coded and age-coded the file. We matched the spousal and dependent SSN's on the file to other records on the Master File of tax documents, and put together families of tax returns. To address issues of changes in taxpayer behavior over time, we included a sub-sample consisting of individuals who were in our Tax Year 1987 sample. By matching to business tax returns, we obtained industry codes for the taxpayers' employers, and by coding the entry in the occupation box or boxes, we generated occupation codes.

In this paper, we describe how the occupation and industry coding of this database was accomplished. We also compare our results to statistics on employment available from the Bureau of Labor Statistics. Previous papers have detailed the information returns match and age and gender coding aspects of the database (Sailer and Weber, 1998), as well as the family linkages it makes possible (Sailer and Weber, 1996 and 1997).

Industry Coding the File

In theory, at least, generating industry codes for a sample of tax return-filing employees is quite easy and inexpensive, given the files at the IRS's disposal. The tax return has SSNs for the primary and secondary taxpayers. These allow the IRS to match the return to the Form W-2 issued by the employer. The W-2, in turn, contains an Employer Identification Number or EIN. The EIN can be used to access the business tax return of the employer, on which an industry code is to be reported. As long as all the returns have been filed, an industry code is just two matches away. Even if the worker has not filed a return, all we need is a Form W-2 that matches to an employer's return record, and we have an industry coded employee.

Actually, the process turned out to be a little more complicated than that. For it was not only the employees who could be non-filers; employers could be non-filers as well. This was pretty unlikely for businesses with employees, but governmental bodies are not required to file tax returns. And while many non-profit organizations do file information returns, these were not on the Business Master File. However, since the employer names were available from the W-2's, we could in many cases generate occupation codes from those names. For example, "Department of" followed by one of the names of the U.S. Government's cabinet agencies definitely indicated a government employee, as did "State of" followed by a State name. And then we had many documents with the words "school" or "college" or "university" in the employer name. These turned out to be a bit of a problem, since public schools and private schools get completely different SIC codes. At the two-digit level, public schools are included with other governmental institutions, whereas private schools have their own SIC code. We decided our tabulations would just show a general education (governmental and non-governmental) category.

The problems enumerated in the previous paragraph were the ones we had figured out before starting the process of industry-coding the file. Once we ran a preliminary table of our data, we realized that 14 percent of all employees who should have been coded (i.e., individuals with salaries and wages, whether they were primary or secondary taxpayers, or non-filers) were listed as "non-codable."

As mentioned previously, the source of the industry codes on the SOI side is, for the most part, the Master File of Business Income Tax Returns. It relies on self-coding by the individuals—generally accountants—who fill out the tax returns. A lot of them appear to be somewhat lacking in imagination and simply don't bother to fill in one of the numbers provided by IRS on the handy list of industry codes right in the filing instructions. When we pulled up the names of the uncoded employers, we saw a regular Who's Who of industrial giants—companies with words like "airline," "petroleum," or "tobacco" right in their names. The obvious solution was to enter SIC codes for these companies. And to code those employer names that did not clearly indicate the industry, we could look up the names in "Moody's On-Line Service," to which the

SOI Division subscribes, and find an industry code there. These corrections brought our non-codable employees down to 4 percent of the file.

A Look at the Industry Data

Table 1 compares the industry distribution from the SOI (tax return) database to data from the Bureau of Labor Statistics (BLS), covering wage-earners by industrial division. In contrast to the SOI data, which were produced by generating industry codes for each of the 112,167 taxpayers and non-filers in our sample, the BLS data came from a survey of employers. In the Current Employment Statistics program, 400,000 non-farm establishments were asked to report on the number of employees on their payrolls. Each establishment was assigned an industry code.

The two major differences between the IRS and the BLS data at the industrial division level are in the manufacturing division (where the IRS data appear high) and in the wholesale division (where the IRS data appear low). These differences may be related to each other. There is a general rule in the IRS Instruction Booklets that tells the company to choose the industry code corresponding to the activity from which it derives the largest percentage of its gross receipts. This code then applies to the whole company, not just (as is true for the BLS data) to a single establishment. Many manufacturing companies are likely to have establishments that engage in wholesaling its products, but these establishments would not be coded separately on the SOI side. In addition, the instruction booklet for Form 1120 (Corporations) specifies that if the company purchases raw materials and sells finished products, it is a manufacturer, even if it contracts out for the labor to make the finished products.

When the data are examined below the industrial division level, such as for the industrial group, the "one code must fit all operations" rule has an even stronger effect. In addition, some industrial group codes do not appear in all the instruction booklets. For example, there is no specific code for engineering and accounting firms on the corporation form, no specific code for petroleum refining on the partnership form, and no specific code for the production of tobacco products on the sole proprietorship form. So the preparer will have no choice but to use the "other" category for such taxpayers. Some of these problems were overcome through the judicious combination of industrial group codes (the first two digits of the 4-digit "Standard Industrial Classification" or SIC code). However, a few groups (notably the miscellaneous manufacturing industry) remain overstated in Table 1.

A few more words about the comparability of SOI and BLS data. Several compromises had to be made when Table 1 was produced. For example, the BLS figures are monthly (based on the pay period that contained the 12th day of each month); the IRS figures, by contrast, are on an annual basis. Since our main goal was to check whether the IRS figures were reasonable, we selected the month with the highest employment figure for the year. On the IRS side, we chose one W-2 for each taxpayer with salaries and wages—the W-2 with the largest salary amount for the year—and used that employer's industry code. So the BLS figures represent the highest employment rate for a given industry, whereas the IRS figures represent, in general, the industries in which taxpayers worked the most. So you would expect the IRS figures to be a little lower than the BLS figures—indeed, the overall IRS figure falls short by 1.92 percent. All in all, given the limitations of the coding methods—one code must fit the whole firm, and not all industrial group codes were available to all taxpayers—we are satisfied that the industry coding worked quite well.

Occupation Coding the File

The tax return offers the U.S. population its only annual opportunity to tell the Federal government what kind of work it is doing. Unfortunately, taxpayers are given very little help in making this report. For industry coding, they may have only one page of codes and one instruction. But for the occupation, they have zero pages of codes and zero instructions—just two boxes that are about 2½ inches long and 1/4 inch tall, labelled "Your occupation" and "Spouse's occupation," respectively, and a gentle reminder: "Don't forget to enter your occupation". SOI gets to decipher what the taxpayers' entries mean.

Now, luckily, we have been doing this for some years, ever since the 1980 Standard Occupational Classification system was devised. For the most part, these have been small studies for subsets of the U.S. taxfiling population, although we did code the full 1979 Statistics of Income sample. Extensive analyses of this projects were presented at various meetings of the American Statistical Association (see Sailer et al., 1980, 1983, 1989, 1990, 1991). Tax Year 1993 marks the first full SOI sample coding effort since then. Of course, we kept all our coding decisions in a computerized dictionary, so any occupation titles coded in previous studies were coded automatically by the computer. (In some cases, it took both the title and an industry code). To help us code new titles that were similar to ones already coded, we hired a contractor to develop a utility similar to a spell-checker—when an uncoded title appeared, it looked for similar word that

had already been coded. All in all, this utility was a great help, although some operators may have been a bit too eager to click the "OK" button. For example, when one taxpayer simply called himself a "professional," the utility helpfully found the code for "Professional Athlete." One simple click of the "OK" button, and all "professionals," no matter what their industry, became athletes. We trust that our subsequent quality review found most of these errors.

From our previous experience, we already knew that, given the level of precision of many taxpayer entries, it would be futile to try to code the file at anything below the two-digit SOC level. Even the two-digit major occupational groups were sometimes too detailed. For example, a frequent taxpayer occupational entry is "nurse;" in order to code it at the two-digit level, we would need to know whether the person was a licensed practical nurse or a registered nurse. Another frequent entry was "operator," which can be coded in conjunction with an industry code. However, to code the individual at the two-digit level, you would need to know whether he was a set-up operator or a production operator. So we ended up consolidating the 60 occupational groups shown in the SOC manual into the 31 groups shown in Table 2. For comparison purposes, we used occupational data for 1993 derived from the Current Population Survey (CPS), a monthly survey by the Census Bureau of 60,000 occupied households. The CPS occupational data are published by the Bureau of Labor Statistics in the series *Employment and Earnings.* As is true of the SOI data, CPS estimates are subject to sampling variability. Since the BLS occupation data (in contrast to the industry data cited earlier) included self-employed individuals, self-employed taxpayers were included on the SOI side as well. However, contrary to what we did with the industry distribution, we could not include non-filers in this tabulation, since we needed a tax return to get an occupation title. We did follow the SOI convention of including late-filed prior-year returns received during 1994, as a stand-in for 1993 returns yet to be filed.

In presenting the SOI occupational data, we decided to create one additional group not part of the SOC coding manual. Because of the vagueness of some titles (most notably, "government worker"), and because an extraordinary number of taxpayers with government industry codes had no occupation entries, we decided to create another category "Government Workers Not Elsewhere Classified."

One more adjustment to the data was needed. The Statistics of Income Division has found that it is not necessary to do an independent edit of Form 1040-EZ for statistical purposes. All the money amounts for these simple returns are already on the IRS Master File of Individual Income Tax Returns, so why not just bring them into the SOI sample unchanged, unless a consistency test shows that the income items are out-of-balance? The plan worked perfectly for all data items except the occupation title. In our database, it is present for the electronically filed Forms 1040-EZ, but not for the corresponding paper forms.

Since the object of this analysis is to evaluate the coverage of various occupations on tax returns, we did not want to simply exclude the filers of paper Forms 1040-EZ. A detailed examination of all Forms 1040-EZ in our sample revealed that all income classes and most industrial divisions represented by paper Form 1040-EZ filers were accounted for among the electronically filed Forms 1040-EZ—although the low-income returns were proportionately underrepresented. Therefore, we weighted up the electronically filed 1040-EZ's to represent all 1040-EZ's. The method we devised controlled both for income size and for industry. By doing this, we made our non-codable records go down from 28 million to 16 million. And while we were doing so, we increased considerably the numbers of transportation, production, and construction workers shown in our tabulations.

A Look at the Occupation Data

Table 2 presents the results of our occupation coding effort. It shows that we succeeded in assigning actual 2-digit SOC codes to 84 percent of the file, with the remainder falling in the "Government Workers Not Elsewhere Classified" or "Non-codable" categories. (Note to all bureaucrats: "Government Worker" is not an occupation.) Not unexpectedly, those occupational groups associated with government work—public officials, social scientists and urban planners, protective service, archivists and curators—are somewhat understated. The only category that is severely overstated is "engineers." At first we thought that a problem we had encountered in the 1979 study—the "building engineer" (who is really a janitor) and the "railroad engineer" (who is really a locomotive operator)—had reappeared. But a careful examination of the occupation titles, employer names, and employer industry codes for everybody coded as an engineer revealed no such obvious problems. One interesting phenomenon we observed was the presence of a fair number or taxpayers who put "Engineer" as their first entry, followed by "Professor" or "Instructor." In the case of multiple entries, we always code to the first entry, on the assumption that it represents the taxpayer's primary concept of his or her job. We could have tweaked the data a little more and brought down

the engineers and raised the college and university teachers a bit—but unless we were going to come up with an alternative coding principal that we could replicate across the board, we did not think that would be the right thing to do.

The understatement of college and university teachers is probably directly related to the overstatement of engineers, architects, and surveyors. There may be some teachers hidden in the data for other professions as well. The understatement of the service occupations, especially private household workers, is probably a true reflection of their under-representation in the tax filing population. The same is probably true of agricultural workers.

Our main objective in this study was to develop procedures that would allow us to occupation-code statistical files relatively quickly and cheaply, using automation as much as possible. It is the authors' opinion that this objective has largely been met. The occupation–coded database should be helpful in a number of ways. Obviously, if anybody wants to do a study of the taxation of the top managers in private industry, of lawyers, of educational counselors, of people in the health diagnosing and treating professions, of technologists, or of mechanics, we can assure them that we have a reasonably good sample of these individuals. If they want to study engineers, they can do that as well, as long as they understand that the sample will include some teaching engineers. Other occupational groupings can still be used, as long as it is clear to the user that they are incomplete. When the Treasury Department builds its Tax Model once every four years or so, it does a statistical match to other files, including the public use file from the CPS. Having good occupation and industry data for over 80 percent of the file will give them two more variables to use for their statistical match, and should improve the quality of their model, even if we haven't coded every last taxpayer.

At this point, it is traditional to say that, of course, much more research is needed. In this case, it is hard to see how much good could come from more research. There are obvious ways improving the occupation data, such as providing more detailed reporting instructions, or asking employers to provide occupation codes on Forms W-2. Because of the additional reporting burden these solutions would impose on taxpayers, they are unlikely to happen. For the foreseeable future, the methods of industry- and occupation-coding described in this paper will be the best that can be done with tax returns.

References

Bureau of Labor Statistics (1999), "Employees on Nonfarm Payrolls by Industry" available on the Internet at http://stats.bls.gov, Table B-1.

Bureau of Labor Statistics (1994), "Employed Persons by Occupation, Sex, and Age," *Employment and Earnings*, Table A-19.

Clark, Bobby; Riley, Dodie; and Sailer, Peter (1989), "1979 Occupation Study/1979-1983 Mortality Study," *Statistics of Income and Related Administrative Record Research: 1988-1989*, Internal Revenue Service, pp. 181-187.

Crabbe, Patricia; Sailer, Peter; and Kilss, Beth (1983), "Occupation Data From Tax Returns: A Progress Report," *Statistics of Income and Related Administrative Record Research: 1983*, Internal Revenue Service, pp. 59-64.

Sailer, Peter; Windheim, Barry; and Fernandez, Mario (1990), Some Results From the 1979-83 Occupational Mortality Study," *Statistics of Income and Related Administrative Record Research: 1990*, Internal Revenue Service, pp. 63-68.

Sailer, Peter; Orcutt, Harriet; and Clark, Phil (1980), "Coming Soon: Taxpayer Data Classified by Occupation," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 467-471.

Sailer, Peter, and Riley, Dodie (1991), "Further Results From the 1979-83 Occupational Mortality Study," *Statistics of Income and Related Administrative Record Research: 1991-1992*, Internal Revenue Service, pp. 39-46.

Sailer, Peter, and Weber, Michael (1996), "Household and Individual Income Data From Tax Returns," *Turning Administrative Systems Into Information Systems:1996-1997*, Internal Revenue Service, pp. 35-41.

Sailer, Peter, and Weber, Michael (1997), "Creating Household Data From Individual Income Tax Returns," *Turning Administrative Systems Into Information Systems:1996-1997*, Internal Revenue Service, pp. 51-55

SOURCE: *Turning Administrative Systems Into Information Systems*, Statistics of Income Division, Internal Revenue Service, as presented at the 1999

Joint Statistical Meetings of the American Statistical Association, Baltimore, MD., August, 1999.