PRELIMINARY ESTIMATES FROM THE 1990 SOI CORPORATE SAMPLE

Susan Hinkins and Jeri Mulrow, Internal Revenue Service

Key Words: Advance data, administrative records

1. INTRODUCTION

The U.S. corporate tax return contains a great deal of economic and tax information that is collected primarily to determine if the calculation of the "tax due" is correct. Therefore, for tax purposes, only a limited amount of information from each return is maintained in a nationwide computer data base at the Internal Revenue Service. The Statistics of Income Division (SOI), within the IRS, produces more comprehensive and accurate data bases by collecting and cleaning essentially all information from samples of these (pre-audit) tax returns, on an annual basis. For a general overview of the SOI missions, programs, and processes, see the paper by Scheuren and Petska.

The tax and income data made available by SOI are important to the U.S. Federal executive and legislative branches as a key source of information for revenue estimation and for analysis of the tax system. The data are used both to analyze the effects of current tax policy and to estimate effects of proposed policy changes. SOI data are also used extensively to measure and analyze the U.S. economy in the National Income and Product Accounts of the U.S. Commerce Department's Bureau of Economic Analysis (BEA). In particular, the SOI corporate data enters into their estimates of GNP.

A common problem with administrative data files is timeliness, and this is a problem with tax data. Consider policy makers interested in economic and tax activity for corporations in 1990. Those tax returns will generally be filed between June 1990 and March 1992. Most of the largest or most complex returns will be filed after September of 1991. The process of retrieving a selected tax return, from the time it is filed, typically takes three weeks, but can take as long as three months. Therefore, the 1990 sample selection takes place between June 1990 and June 1992. In addition to this is the time needed to abstract the data, check and clean the data file, and calculate estimates. Under the old SOI processing system, the final estimates for 1990 would not be ready until March 1993.

However, the data users need information and estimates before that time. And in fact the ultimate goal is to provide estimates "on demand," before the sampling is complete. As a starting point, we are providing advance data estimates and an advance data file at one particular time point. For the 1990 sample, the advance data were provided by May 1, 1992.

This requires three new processes. First, a new data entry system was needed that shortens the time to enter and check the data. Second, because the "late" returns are not like the "early" returns, the properties of the missing returns need to be modeled. Finally, there are critical corporations that must be in the data base in order to make reasonable estimates. A means for identifying these corporations, tracking them, and getting that information was needed.

In Section 2, we briefly describe the new data processing system that made advance data estimation feasible. In Section 3 we discuss the very simple models used in the first year, and why more sophisticated modeling was put-off until the 1991 sample. Section 4 describes the "critical" corporations, and the use of a survey to add their information to the advance data.

2. IMPROVED DATA ENTRY SYSTEM

The estimation process is highly dependent on the data collection process. Advance estimates were not even feasible under the old process of entering and checking the data. It only became feasible for the 1990 sample when having more timely data became so important to one of our primary users, BEA, that they provided funding for a new processing system.

Under the old system, data from the tax return were first entered onto (paper) edit sheets. The people doing the data entering are referred to as editors, because it is not simply data abstraction. Tax return data are relatively clean and correct, but errors can exist. A correct amount may be entered on the wrong line, or in some cases, the economic variable of interest is slightly different than the tax line. Industrial classifications are not always entered by the taxpayer and must be determined by the editor, etc.

From the edit sheets the data were keyed into the computer file. These records, in batches, were then run through a series of consistency tests; any errors were printed out and sent back to the editors. Typically, a different editor would then go back to the tax return, at least a week after the initial data editing, and attempt to correct the record. Changes would be noted and another person would key the changes. The record would go back through the consistency tests, and this cycle would continue until the record was corrected, or time ran out.

The inefficiencies and disadvantages of this system are obvious. SOI had wanted to move to an interactive data entry and testing system for many years. The stumbling block was purely resources. These resources were provided by BEA for the 1990 system. Under the new system, the data are entered and tested at one sitting. An interactive system checks the data as they are entered and prompts the editor to correct problems or look for missing entries. The advantages of this system are obvious. The quality of the data should be better and the editing process should be easier and faster. In particular, there will be more clean, complete records available for analysis at an earlier time.

This new processing system allows final estimates to be made at an earlier time and makes advance estimation feasible.

3. MODELING "MISSING" RETURNS

There are two major statistical problems associated with these advance estimates from incomplete data. First, the missing records, the "late" returns, have different properties than the "early" returns. Therefore, modeling of these returns is necessary.

Second, the distribution of most of the variables of interest is highly skewed. That is, a relatively few number of units account for a large percentage of the total amount. For example, in 1989 the largest 0.02 percent of the corporations contained 53 percent of the total (national) amount in the variable "Interest Received," and 80 percent of the total amount in "Foreign Dividend Gross-up." Therefore, there are certain units that are extremely influential to the estimates. Unfortunately, these influential observations are more likely to be "late."

Modeling is the more interesting statistical problem; books and papers are written on models. The need for defining 100 percent strata can be discussed in a paragraph in most sampling texts. However, in the corporate population, the very largest corporations are so influential to the estimates that almost all our time in the first year was devoted to capturing these data, as described in Section 4. These records have to be in the advance data base in order to make reasonable estimates and without these data, modeling efforts are wasted.

The modeling effort for the 1990 advance estimates was minimal. Estimates of the final population size, N, in each sampling strata were made, and the advance data sample sizes, m, were known for each stratum. Weighted estimates were calculated using N/m as the weight, by sampling strata. Records designated as critical were given a weight of one.

Because the late records have different properties than the advance sample, we knew that treating the advance data as a random subsample of the final sample in this way is not adequate, but may result in biased estimates. Therefore, for the variables designated by BEA as most important, ratio adjustments were made to the weighted estimates. Using prior year data, the ratios of the final estimate to the advance data estimate were calculated, by industrial classifications. These ratios were relatively stable from 1988 to 1989. Two estimates were provided to the users: first adjusting by the 1989 ratio and second, adjusting by an average of the 1989 and 1988 results. We will be evaluating how well this simple model worked as soon as the final file is complete.

4. CRITICAL CASES

Finally, because of the extremely skewed distribution of many of the economic and tax variables, it was determined that the most important task in the first year was to ensure that these critical corporations were in the advance data. These largest corporations are so influential and, at least for certain variables, so unstable from year to year that modeling these records for the advance data is not a reasonable option, as will be shown.

The effect of the skewed distribution can be seen in the overall sample design. There are nearly 4 million corporate returns in an annual population. The size of the sample is usually between 75,000 and 85,000, and almost one fourth of the sample is devoted to 100 percent strata, i.e., records that are sampled with probability one.

For the advance data, we needed to designate a reasonable number of the largest corporations. The primary user of the advance data, BEA, selected a definition of critical corporations, using the variable "Total Assets," that resulted in a list of approximately 675 corporations, based on 1988 and 1989 records.

To make sure that these data were in the advance sample, a small survey was added to the administrative data base. For these critical corporations, where the return was not going to be in the advance sample, a short questionnaire was sent directly to the corporation requesting approximately 15 tax items that were considered most important by our user. In this way at least some of the current information for these corporations was obtained. The response to this survey was surprisingly good.

Of the 675 corporations designated as critical, 650 were still in the population in 1990. Twenty-two had filed as a subsidiary of another corporation in 1990; that is, these corporations are still represented in the population but as part of another corporation. Three had no filing requirement in 1990.

At the time of the advance data file, out of the 650 critical cases, only 3 corporations had no 1990 data. These were corporations that were not in the sample in time and did not respond to the (voluntary) survey. There were only 19 corporations for which we had only the survey data; that is, 19 corporations that were not in the advance sample but did respond to the survey.

The process of tracking the corporations was a lot of work: determining if they were in the 1990 population, if they would file in time for the advance data, sending out questionnaires, following up, etc. Was it worth all the effort to collect the current data for these 19 units? Would it have made a noticeable difference to the estimates if we had simply used the prior year information for these records? To answer these questions, we compare the estimates using the survey data for the 19 records to the estimates calculated by replacing the 1990 survey data with the 1989 data. Only those 15-20 variables collected on the short survey are considered here.

If the user is only interested in the population totals, then these 19 records would not be critical. The largest difference in estimated totals between using prior year data vs. the survey data is $10 billion. While this seems like a significant difference, it represents only 0.1 percent difference in the estimate of "Total Receipts." The largest percentage difference is - 0.5 percent, in "Dividends."

However, advance data estimates are also needed for subpopulations defined by industrial categories. The largest subpopulations are 10 industrial divisions, subdividing the population into Agriculture, Mining, Finance, Construction, etc. These are large categories and one would not expect that 19 records would seriously effect estimates at this level, unless all or most of the missing records were in one division. This did not happen. The 19 records were spread over 7 out of 10 divisions.

Table 1 shows the most extreme differences in division estimates that would have occurred if prior year data had been used instead of the 1990 survey data, for just two corporations. If for just those two critical records, we had used the 1989 values for "Total Dividends," the <u>division</u> estimate of "Total Dividends" would be 26 percent smaller than the estimate using the 1990 values for these two records.

The users also need estimates for even smaller subpopulations. The population can be further divided into 58 major industrial classes. For example, the Finance division is divided into categories such as Banking, Insurance, Real Estate, etc. Given the results in Table 1 for one division, we know that there is at least one example where the survey data make a noticeable difference in major industry estimates. There are in fact two major industries with extreme results. Only one is contained in the division of Table 1. In each case, there is only one critical record used in the estimation.

The second half of Table 1 shows the effects of one critical record on estimates for one of the major indus-

Table 1.--Using Prior Year Data Compared to Survey Data

In One Industrial Division
(two records with survey data)

The Division estimate of	Would have been
Cash & Property Distributions	4% smaller
Net Gain	8% larger
Total Dividends	26% smaller

For a Major Industrial Class
(one record with survey data)

The estimate of	Would have been
Interest Received	15% larger
Taxes After Credits	17% smaller
Cash & Property Distributions	23% smaller
Net Gain	54% larger
Total Dividends	76% smaller

tries contained in this division. Estimates of five variables would have changed by more than 10 percent. For example, using the prior year value for "Net Gain" in just one record would have resulted in an industry estimate 54 percent larger than the estimate using the current year value.

There are also dramatic differences in two estimates in one other major industry, in a different division. In this case, both variables are associated with gain or loss, which are often unstable variables. For this one critical corporation, the 1990 value of "Net Gain" is 16 times the 1989 value. Using the prior year value would result in an industry estimate of "Net Gain" that is 15 percent smaller than if the current year value is used. The variable "Income or Loss from Foreign Sources" can be positive or negative, so that relative changes can be quite large. In particular, for this corporation it went froma loss in 1989 to a gain in 1990 of almost twice the magnitude. Using the 1989 value would have resulted in an industry estimate 91 percent smaller than the estimate using the current value.

5. CONCLUSIONS

Two conclusions seem apparent. First, these corporations designated as critical are not misnamed. They are extremely influential for subpopulation estimates, even for quite large or general subpopulations. Since our users are often interested in smaller subpopulations than shown here, the effect of these corporations can be even greater. Without these units in the sample, the error bounds on the estimates would be so large as to make the subpopulation estimates useless.

Second, the year-to-year variability in some variables, at the record level, can be quite large, in both absolute and relative terms. These largest corporations are so influential and, at least for certain variables, so unstable that modeling these records for the advance data does not seem to be a reasonable option.

The future plans encompass many improvements to this process. The definition of critical corporations is being evaluated and the survey methods improved. There is much to be done in modeling the missing corporations, and in particular, we are looking at estimating the propensity to be filed by a given time, to be used in weighting the records, plus using ratio adjustments for the estimates of ttotals. By improving the data entry process, using more modeling techniques, and adding a small survey to the administrative data base, we expect to produce much more useful information for our users by providing a more timely data base.

REFERENCE

Scheuren, F. and Petska, T., "Turning Administrative Systems into Information Systems," to appear in the *Journal of Official Statistics*, Statistics Sweden.