Section 3 Description of the Sample and Limitations of the Data

This section describes the 1995 Corporate sample design, including the methods used in the selection of returns, data capture, data cleaning, and data completion. The techniques used to produce estimates and an assessment of the data limitations, including measures of sampling variability, are also discussed.

Background

From Tax Year 1916 through Tax Year 1950, data were extracted for the Statistics of Income (SOI) program from each corporate return filed. Stratified probability sampling was introduced for Tax Year 1951. Since then, the size of the samples has generally decreased while the population has increased. For example, for Tax Year 1951 the sample comprised 41.5 percent of the entire population, or 285,000 of the 687,000 total returns filed. For 1995, the sample proportion had decreased to about 2.0 percent from a total population of over 4.8 million.

For 1951, stratification was by size of total assets and industry. From 1952 through 1967, the stratification was by size only. The size was measured by volume of business (1953-1958) or total assets (1952, and 1959-1967). Since 1968, returns have been stratified by both total assets and a measure of income, the definition of which depends on the return's form type [1].

Population

Statistics in this report are for active corporations organized for profit and required to file the Form 1120 series returns. The following table gives the actual number of corporate returns by form type that were filed during Tax Years 1991 through 1995 and were, therefore, subject to sampling. These population counts will differ from all the estimated population counts in this publication because they include out-of-scope returns which are excluded from the tabulations (see page 10).

Bertrand Überall, Susan Hinkins, Richard Collins, and Valerie Puckett were responsible for the sample design and estimation of the SOI 1995 Corporation Program under the direction of Yahia Ahmed, Chief, Mathematical Statistics Section, Statistical Computing Branch.

Form Type	Tax Year 1991	1992	1993	1994	1995
1120	2,013,142	1,951,065	1,980,483	2,214,657	2,235,287
1120-A	333,593	355,330	325,773	321,402	325,249
1120S	1,763,665	1,875,837	2,011,167	2,139,353	2,267,178
1120-L	2,098	2,020	1,942	1,829	1,718
1120-PC	2,624	2,710	2,760	2,846	2,928
1120-RIC	5,368	6,038	6,931	7,712	8,478
1120-REIT	283	290	354	394	473
1120-F	12,222	11,544	11,274	11,905	10,875
Total	4,132,995	4,204,834	4,340,684	4,700,098	4,852,186

Sample Design

The current sample design is a stratified probability sample, with stratification by form type, and either size of total assets alone, or both size of total assets and a measure of income. Forms 1120 and 1120-A are stratified by size of total assets and size of "proceeds." Size of "proceeds" is used as the measure of income, and is defined to be the larger of the absolute value of net income (or deficit) or the absolute value of "cash flow," which is the sum of net income and several depreciation amounts. Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, and 1120-REIT are each stratified by size of total assets only. Form 1120S is stratified by size of total assets and, as the measure of income, size of ordinary income.

The design process began with projected population totals derived from those used to estimate IRS administrative workloads and are adjusted based on previous years' population distributions. Using projected population totals by sampling strata, an optimal allocation, based on variance and cost estimates, was carried out to assign sample rates such that the overall projected sample size is 92,000. A Bernoulli sample is selected independently from each stratum with rates ranging from .25 percent to 100 percent. The selected sample size for Tax Year 1995, including inactive corporations and rejected returns, is 97,605 returns for all form types. Figure C on the following page shows the stratum boundaries, population and sample sizes, as well as the adjusted population and sample sizes for all form types after reclassification of returns due to errors in the stratifying variables (see subsection on Nonsampling Error, page 13, for further information on the handling of mis-stratified returns) and the achieved sample rates after reclassification.

Figure C--Corporation Returns: Number Filed, Number in Sample, and Sampling Rates by Sample Selection Class

Sample Class Number	Description of Sample Selection Classes		Sampling Rates (%)	Number of Returns			
	Size of Total Assets (000's)	*Size of Proceeds (000's)		Original BMF		After Re-stratification	
				Population	Sample	**Population	***Sample
	All Returns, Total			**4,852,186**	**97,605**	**4,852,305**	**97,461**
	Forms 1120 w/ Form 5735 Attached, Total			**496**	**496**	**493**	**493**
1	Under $100,000		100.00	403	403	401	401
2	$100,000 - $250,000		100.00	44	44	44	44
3	$250,000 or more		100.00	49	49	48	48
	Forms 1120 (no Form 5735 attached), 1120-A, Total			**2,560,040**	**62,902**	**2,560,125**	**62,827**
4	Under $50	Under $25	0.48	986,949	4,724	984,343	4,748
5	$50 - $100	$25 - $50	0.48	359,062	1,681	359,204	1,715
6	$100 - $250	$50 - $100	0.74	454,606	3,449	455,510	3,509
7	$250 - $500	$100 - $250	1.54	290,306	4,473	291,890	4,533
8	$500 - $1,000	$250 - $500	2.50	193,329	4,790	193,900	4,870
9	$1,000 - $2,500	$500 - $1,000	5.00	142,914	7,087	143,564	7,151
10	$2,500 - $5,000	$1,000 - $1,500	6.20	53,443	3,348	53,079	3,368
11	$5,000 - $10,000	$1,500 - $2,500	11.00	29,391	3,108	29,117	3,105
12	$10,000 - $25,000	$2,500 - $5,000	30.00	20,748	6,114	20,521	6,035
13	$25,000 - $50,000	$5,000 - $10,000	50.00	10,368	5,204	10,245	5,113
14	$50,000 - $100,000	$10,000 - $15,000	100.00	6,888	6,888	6,814	6,776
15	$100,000 - $250,000	$15,000 or more	100.00	6,448	6,448	6,367	6,333
16	$250,000 or more		100.00	5,588	5,588	5,571	5,571
	Forms 1120S, Total			**2,267,178**	**24,192**	**2,267,203**	**24,145**
17	Under $50	Under $25	0.25	976,662	2,389	967,928	2,415
18	$50 - $100	$25 - $50	0.32	356,183	1,142	356,258	1,181
19	$100 - $250	$50 - $100	0.48	382,015	1,875	386,157	1,934
20	$250 - $500	$100 - $250	1.00	236,262	2,400	240,344	2,453
21	$500 - $1,000.	$250 - $500	1.80	137,083	2,454	137,905	2,505
22	$1,000 - $2,500	$500 - $1,000.	4.00	100,067	3,997	100,478	4,025
23	$2,500 - $5,000	$1,000 - $1,500	5.00	40,410	2,008	40,363	2,022
24	$5,000 - $10,000	$1,500 - $2,500	9.60	21,837	2,065	21,631	2,059
25	$10,000 - $25,000	$2,500 - $5,000	24.00	11,702	2,766	11,505	2,718
26	$25,000 - $50,000	$5,000 - $10,000	40.00	3,154	1,293	3,044	1,245
27	$50,000 - $100,000	$10,000 - $15,000	100.00	1,021	1,021	959	957
28	$100,000 - $250,000	$15,000 or more	100.00	621	621	554	554
29	$250,000 or more		100.00	161	161	77	77
	Forms 1120-L, Total			**1,718**	**957**	**1,706**	**944**
30	Under $50,000		43.00	1,363	602	1,324	562
31	$50,000 - $250,000.		100.00	119	119	127	127
32	$250,000 or more		100.00	236	236	255	255
	Forms 1120-F (with effectively-connected income in U.S.), Total			**10,875**	**1,787**	**10,868**	**1,774**
33	Under $50,000		14.00	10,587	1,499	10,579	1,486
34	$50,000 - $100,000.		100.00	68	68	68	67
35	$100,000 or more		100.00	220	220	221	221
	Forms 1120-PC, Total.			**2,928**	**1,122**	**2,960**	**1,134**
36	Under $100.		30.00	601	191	492	144
37	$100 - $5,000		30.00	1,315	386	1,323	390
38	$5,000 - $50,000.		30.00	684	217	792	247
39	$50,000 - $250,000		100.00	208	208	217	217
40	$250,000 or more		100.00	120	120	136	136
	Forms 1120-REIT, Total			**473**	**370**	**476**	**368**
41	Under $50,000		50.00	207	104	209	105
42	$50,000 - $250,000.		100.00	133	133	134	130
43	$250,000 or more		100.00	133	133	133	133
	Forms 1120-RIC, Total			**8,478**	**5,779**	**8,474**	**5,776**
44	Under $50,000		25.00	3,568	869	3,554	856
45	$50,000 - $250,000.		100.00	2,705	2,705	2,713	2,713
46	$250,000 or more		100.00	2,205	2,205	2,207	2,207

* Proceeds is defined as the larger of absolute value of net income (deficit) or absolute value of cash flow (depreciation + depletion + net income).

** Includes added returns not processed in the BMF during the two-year sampling frame.

*** Does not include missing returns.

Note: Returns were classified according to either size of total assets or size of proceeds, whichever corresponded to the higher sample class.
EXAMPLE: A Form 1120 return with total assets of $750,000 and a proceeds of $75,000 is in sample class 8 (based on total assets), rather than in sample class 6 (based on proceeds).

Sample Selection

Corporation income tax returns are filed at the ten IRS service centers located throughout the country. All corporate returns are processed initially to determine tax liability and are then made available for other programs including SOI. All tax data are transmitted and updated on a weekly basis to the IRS Business Master File (BMF) system located in Martinsburg, West Virginia. This system serves as the point of selection for the sample, which was selected on a weekly basis.

Sample selections for Tax Year 1995 occurred over the period of July 1995 through June 1997. A 24-month sampling period is needed for two reasons. First, approximately 24.2 percent of all corporations have noncalendar year accounting periods. In order to take the noncalendar filings into consideration, the 1995 statistics represent all corporations filing returns with accounting periods ending during the period from July 1995 to June 1996. Also, many corporations, including some of the largest, request 6-month filing extensions. The combination of noncalendar year filing and filing extensions means that the last returns due to be received by IRS for the Tax Year 1995 (those with accounting periods ending in June 1996, which must therefore be filed by October 1996) could be timely filed as late as March 1997, if the 6-month extension of the October 1996 due date is taken into account. Normal administrative processing time lags required that the sampling process remain open for the 1995 study until June 30, 1997. However, a few very large returns for Tax Year 1995 were added to the sample as late as November 1997

Each corporation is assigned a permanent and unique Employer Identification Number (EIN). The EIN is used as the basis for random selection. A pseudo-random number (PRN) is generated using the EIN as the seed. The last four digits of the PRN, called the transformed taxpayer identification number (TTIN), are compared to the sampling rates; a corporation for which the value of its TTIN is below the sampling rate multiplied by 10,000 is selected in the sample. The algorithm for generating the TTIN does not change from year to year. Consequently, any corporation selected into the sample in a given year will be selected again the next year, providing that the corporation files a return using the same EIN in the two years and that it falls into a stratum with the same or higher rate. If the corporation falls into a stratum with a lower rate, the chance of selection will correspond to the ratio of the second year to the first year selection probabilities. If the corporation files with a new EIN, the probability of being selected will be independent of the prior year selection. Due to the fact that corporations typically maintain the same EINs, this use of the EIN for the basis of sample selection results in many of the same corporations selected into the samples from year to year. This also results in a reduction of the sample variance for estimates of year-to-year change [2].

Data Capture

Data processing for SOI begins with information already extracted for administrative purposes; over 100 items are available from the BMF system for nonconsolidated Form 1120 returns. Some 900 additional items are extracted from the tax returns. The administrative data are checked and corrected as necessary. The SOI data capture process can take as little time as fifteen minutes for a small, single entity corporation filing on Form 1120-A, or as long as a week for a large consolidated corporation filing several hundred attachments and schedules with the return. The process is further complicated by several factors:

¦ The 900 separate data items that may be extracted from any given tax return often require totals to be constructed from various other items on other parts of the return.

¦ Each 1120 form type has a different layout with different types of schedules and attachments, making data extraction less than uniform for the various form types.

¦ There is no legal requirement that a corporation meet its tax return filing requirements by filling in, line for line, the entire U.S. tax return form. Therefore, many corporate taxpayers report many of their financial details in schedules of their own design.

¦ There is no single accepted method of corporate accounting used throughout the country, but rather several accepted accounting "guidelines," many of which are unique to geographic locations. SOI attempts to standardize these differences during data abstraction and editing.

¦ Different companies may report the same data item, such as other current liabilities, on different lines of the tax form. Again, SOI attempts to standardize these differences.

In order to help overcome these complexities and differences due to taxpayer reporting, SOI prepares detailed instructions for the SOI editing unit at designated service centers each tax year. For Tax Year 1995, these instructions consisted of more than 800 pages covering normal and straightforward procedures and instructions for exceptions and nonstandard situations that might be encountered.

Data Cleaning

Statistical processing of the corporate returns took place in an online computer environment. This means that the data from returns were entered directly into the corporation database. In this context, the term "editing" refers to the combined interactive processes of data extraction, consistency testing, and error resolution. There are over

800 of these tests, which look for such inconsistencies as:

- Impossible conditions, such as incorrect tax data for a particular form type;
- Internal inconsistencies, such as items not adding to totals;
- Questionable values, such as a bank with an unusually large amount reported for cost of goods sold and/or operations; and
- Improper sample class codes, such as when a return has $10,000 in total assets, but was selected as though it had $1 million.

Data Completion

In addition to the tests mentioned above, missing data problems must be addressed and returns that are to be excluded from the tabulations must be identified. The data completion process focuses on these issues.

If the missing data items are from the balance sheet, then imputation procedures are used. If data for a whole return are missing because the return is unavailable to SOI during the data capture process, then, again, imputation procedures are used in certain cases.

A ratio-based imputation procedure is used to estimate missing balance sheet items for all 1120 forms except those with less than 12-month accounting periods. The ratios are determined by the corporation's 1994 return if it is available; otherwise, the 1993 aggregate data for the corporation's minor industrial group are used. If the reported items in the balance sheet do not balance (i.e., the sum of asset items does not equal the sum of liability and shareholders' equity items), then missing items are imputed. If the total assets amount is among the missing items, this item is imputed first based on the ratio of total assets to business receipts (or total receipts) from either the corporation's 1994 return, or the 1993 aggregate data for the corporation's minor industry. The other missing asset and liability items are then imputed based on the ratios so that the total of all asset items and the total of all liability items are both equal to the total assets amount, whether this amount was reported or imputed. For a detailed description of the balance sheet imputation process, see reference [3]. The following table shows the number of sampled returns that had balance sheet items imputed for Tax Years 1991 through 1995.

Tax Year	1991	1992	1993	1994	1995
No. of Returns	303	221	214	230*	131*

* Starting in Tax Year 1994, 1504(c) returns are counted as one return rather than separate entities when computing the number of imputed balance sheets.

For Tax Year 1995, of the 131 returns, 25 of them have imputed total assets, and the imputed total asset amount constitutes approximately .0016 percent of the estimated total assets of the active corporations in 1995.

Data for unavailable critical corporations are imputed in various ways, depending on what information is available at the time the SOI database is produced. Critical corporations include corporations with total assets greater than or equal to 5 percent of the total assets for the minor industrial group in which they are classified, and corporations for which total assets are over a specified limit which is dependent on the form type or the major industry. For critical corporations selected for the sample but unavailable for statistical processing, taxpayer-surveyed data are used. There are two such returns in the Tax Year 1995 data. For the critical corporations not selected for the sample, if the current tax return is not found in any of the IRS service centers and no other current tax data are available, data from the previous year's return are used with adjustments for tax law changes. There is one prior year return in the Tax Year 1995 data.

Another part of the data-cleaning process is identifying sampled returns that are not used in the tabulation. The BMF system, used for sample selection, can include duplicate tax returns and other out-of-scope returns, such as returns for nonprofit corporations and prior-year tax returns. These include the following types of returns:

- Inactive corporation returns (having neither current income nor deductions);
- Duplicate returns;
- Amended returns not associated with the original returns;
- Tentative returns not associated with the revised returns;
- Corporations exempt under Code section 931;
- Corporations exempt under Code section 1247;
- Corporations exempt under Section 883 of the IRC;
- "Cost corporation" returns exempt under Revenue Ruling 52-542;
- Corporations exempt under Code section 501(c)(15);
- Nonresident foreign corporations having no income effectively connected with a trade or business within the United States;
- U.S. Virgin Island corporations exempt under Code section 934;

¦ Political organizations filing under Code section 527;

¦ General stock ownership corporations exempt from tax;

¦ Homeowners' associations under Code section 528;

¦ Information returns reporting no tax because of tax treaty or convention according to Code section 894;

¦ Most prior-year returns with total assets under $250 million filed on tax forms for years prior to 1994 and with accounting periods ending before July 1995;

¦ Returns filed on a form type which should not be included in the SOI sample;

¦ Fraudulent Returns;

¦ Tax-Exempt U.S. Possessions.

The following table displays the number of sampled returns that were excluded from tabulations and the percentages they represent of the total sample sizes in Tax Years 1991 through 1995.

Type of Return	Tax Year 1991	1992	1993	1994	1995
Inactive	978	903	1,188	1,367	1,466
Duplicate	883	213	166	634	984
Other*	1,099	2,213	2,958	2,009	2,217
Total	2,960	3,329	4,312	4,010	4,667
% of Sample	3.50	3.96	4.71	4.22	4.78

* Includes prior-year returns.

The estimated number of active corporations by form type for Tax Years 1991 through 1995 are provided in the next table.

Form Type	Tax Year 1991	1992	1993	1994	1995
1120	1,821,503	1,782,933	1,775,931	2,038,870	2,043,818
1120-A	262,648	280,697	265,627	257,125	257,439
1120S	1,698,271	1,785,371	1,901,505	2,023,754	2,153,119
1120-L	2,026	1,959	1,876	1,775	1,646
1120-PC	2,506	2,475	2,623	2,674	2,789
1120-RIC	5,585	5,842	6,796	7,519	8,201
1120-REIT	291	293	346	393	465
1120-F	9,958	9,452	9,925	10,259	*6,690
Total	3,802,788	3,869,022	3,964,629	4,342,368	4,474,167

Note: Detail may not add to total due to rounding.
* This estimate is significantly lower than in previous years (see section on Coverage Errors).

Estimation

The estimates produced in this report of the total number of corporations and associated money amounts are based on weighted sample results. Either a one-step process or a two-step process was used to determine the weights, depending on the return's form type.

Under the one-step process, the weights are assigned as the reciprocal of the achieved sample rate. These weights are used to produce the aggregated total frequencies and money amounts published in this report for Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, 1120-REIT and Form 1120 with Form 5735 attached.

The two-step process was used to improve the industry estimates. The first stage is identical to the one-step process as described above and provides an initial weight for the return. The second stage involves poststratification by industry. During poststratification, certain cells have small sample sizes. To handle this problem, a raking ratio estimation approach is applied during poststratification in order to determine the final weights [4]. Restrictions are placed on the raking process to produce final weights that fall within the range ***l*** (2/3) x original weight to ***l*** (3/2) x original weight. These final weights are used to produce the aggregated frequencies and money amounts published in this report for Forms 1120, 1120-A and 1120S.

Data Limitations and Measures of Variability

Several extensive quality review processes were used to improve the quality of the data. The review processes began at the sample selection stage with weekly monitoring of the sample to ensure that the proper number of returns was being selected. They continued through the data collection, data cleaning, and data completion procedures with consistency testing. Part of the review process included extensive comparisons between the 1995 data and the 1994 data. A great amount of effort was made at every stage of processing to ensure data integrity.

Sampling Error

Since the corporation estimates are based on a sample, they may differ from figures that would have been obtained if a complete census of all income tax returns had been taken. The particular sample used to produce the results in this report is one of a large number of possible samples that could have been selected under the same sample design. Estimates derived from one of the possible samples could differ from those derived for any other sample, and from the population aggregates. The deviation of a sample estimate from the average of all possible similarly selected samples is called the sampling error. The standard error (SE) is a measure of the average magnitude of the sampling errors over all possible samples.

The standard error is the most commonly used measure of the sampling error and can be estimated from the sample. Sometimes, for convenience, the standard error is expressed as a percentage of the value being estimated. This is called the coefficient of variation (CV) of the estimate. The coefficient of variation can be used in assessing the reliability of an estimate.

The coefficient of variation of an estimate is calculated by dividing the standard error by the estimate. Coefficients of variation by industrial groupings for the estimated number of returns, as well as for selected money amount estimates, are shown in Table 1 beginning on page 29. For the estimated number of returns by asset size and industrial division, coefficients of variation are given in Figure D.

The coefficient of variation, CV(X), can be used to construct confidence intervals of the estimate X. The standard error, which is required for the confidence interval, must first be calculated. For example, the estimated number of manufacturing companies with net income and its coefficient of variation can be found in Table 1 and used to calculate the standard error:

$$\begin{aligned} SE(X) &= X \times CV(X) \\ &= 185{,}590 \times 2.31/100 \\ &= 4{,}287 \end{aligned}$$

Assume that a 95-percent confidence interval for the number of returns in manufacturing is desired. The 95-percent confidence interval is constructed as follows:

$$\begin{aligned} X \pm (2 \times SE(X)) &= 185{,}590 \pm (2 \times 4{,}287) \\ &= 185{,}590 \pm 8{,}574 \end{aligned}$$

Thus, the interval estimate is 177,016 returns to 194,164 returns. This means that if all possible samples were selected under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors

Figure D--CVs for Number of Returns, by Asset Size and Industrial Division, Tax Year 1995

Industrial division	All asset sizes	Size of total assets: Zero assets	$1 under $ 100,000	$100,000 under $250,000	$250,000 under $500,000	$500,000 under $1,000,000
	(1)	(2)	(3)	(4)	(5)	(6)
All Industries[1].	0.19	3.03	0.50	0.81	0.71	0.54
Agriculture.	2.47	17.60	5.84	4.89	3.15	2.70
Mining.	6.28	29.16	13.08	14.80	12.31	9.76
Construction	1.00	10.56	1.99	3.02	2.64	1.87
Manufacturing	1.66	11.92	4.10	4.44	3.25	2.57
Transportation	2.41	15.77	4.09	6.11	5.15	4.75
Wholesale and retail trade	0.65	6.17	1.46	1.61	1.44	1.15
Finance, insurance, and real estate. . .	0.89	6.86	2.01	2.26	1.79	1.46
Services	0.62	5.33	0.96	1.93	2.15	2.02

Industrial division	Size of total assets--continued: $1,000,000 under $5,000,000	$5,000,000 under $10,000,000	$10,000,000 under $25,000,000	$25,000,000 under $50,000,000	$50,000,000 under $100,000,000	$100,000,000 under $250,000,000
	(7)	(8)	(9)	(10)	(11)	(12)
All Industries[1].	0.26	0.59	0.46	0.58	0.04	0.04
Agriculture	2.06	7.17	5.97	5.26	0.92	1.04
Mining.	5.89	7.64	5.64	6.74	0.65	0.70
Construction.	1.09	2.12	2.00	3.14	0.51	0.79
Manufacturing	1.18	1.91	1.18	1.30	0.18	0.18
Transportation	2.50	4.56	2.90	3.53	0.39	0.43
Wholesale and retail trade	0.68	1.25	0.95	1.37	0.21	0.25
Finance, insurance, and real estate. . .	0.89	1.87	1.22	1.18	0.07	0.07
Services	1.33	3.06	2.16	2.37	0.29	0.33

[1]Includes returns not allocable by industrial division.

Note: Returns with total assets of $250 million or more were sampled at the 100% rate and are, therefore, not subjected to sampling error.

above the estimate would include the average estimate derived from all possible samples. Thus, for a particular sample, it can be said with 95-percent confidence that the average of all possible samples is included in the constructed interval. This average of the estimates derived from all possible samples would be equal to or near the value obtained from a census.

Nonsampling Error

In addition to sampling error, nonsampling error can also affect the estimates. Nonsampling errors can be classified into two groups: random errors whose effects may cancel out and systematic errors whose effects tend to remain somewhat fixed and result in bias.

Nonsampling errors can be categorized as coverage errors, nonresponse errors, processing errors, or response errors. These errors can be the result of the inability to obtain information about all returns in the sample, differing interpretations of tax concepts or instructions by the taxpayer, inability of a corporation to provide accurate information at the time of filing (data are collected before auditing), inability to obtain all tax schedules and attachments, errors in recording or coding the data, errors in collecting or cleaning the data, errors made in estimating for missing data, and failure to represent all population units.

Coverage Errors

Coverage errors in the SOI corporation data can result from the difference between the time frame for sampling and the actual time needed for filing and processing the returns. As stated above, many of the largest corporations receive extensions to their filing periods and, as a result, may file their returns after sample selection has ended for that tax year. However, any of the largest returns found are added into the file until the final file is produced.

Coverage problems within industrial divisions in the SOI corporation study result from the way consolidated returns may be filed. The Internal Revenue Code permits a parent corporation to file a single return, which includes the combined financial data of the parent and all its subsidiaries. These data are not separated into the different industries but are entered only into the industry with the largest receipts. Thus, there is undercoverage of financial data within certain industries and overcoverage in others. Coverage problems within industrial divisions present a limitation on any analysis done with the sample results.

In 1995, there was a processing problem prior to the sampling operation which resulted in many 1120-F returns filed by corporations with income “effectively connected with a U.S. trade or business” being incorrectly excluded from the sampling frame. This resulted in undercoverage of the 1120-F population. Specifically the estimated active population (6,690) is significantly lower than that of previous years.

A set of ad hoc adjusted weights gave an estimate of the active population as 9,323. However, other adjustment techniques that we are pursuing may yield a higher or lower estimate. Although the particular set of weights we used did improve the population count, it did not improve the quantitative variable estimates (money amounts), because the missing returns have a different characteristic than the sampled returns. Therefore, it was decided not to include the following two tables in Publication 16 (Corporation Income Tax Returns) for Tax Year 1995: “Income Statement and Selected Tax Items by Industrial Division and Selected Major Industries (Returns of Active Corporations, Form 1120-F)” and “Income Statement and Selected Tax Items by Industrial Division and Selected Major Industries (Returns with Net Income, Form 1120-F),”
Table 10 and Table 11, respectively, in Publication 16 for Tax Year 1994. Currently, SOI is researching different methods to adjust the quantitative variable estimates. The results will be published in a future issue of the *SOI Bulletin.*

Nonresponse Errors

Unit nonresponse for SOI occurs when a sampled return is unavailable for SOI processing. For example, other areas of the IRS such as Examination, Collection, or District Offices may have the return at the time the return is needed for statistical processing. These returns are termed "unavailable returns." In 1995, there were 138 unavailable returns in the corporation study, which constituted about .14 percent of the total sample size. The following table gives the number of unavailable returns and their percentages of total sample sizes for Tax Years 1991 through 1995.

Tax Year	1991	1992	1993	1994	1995
No. of Returns	69	99	118	113	138
% of Sample	0.08	0.12	0.13	0.12	0.14

Processing Errors

Errors in recording, coding or processing the data can cause a return to be sampled in the wrong sampling class. This type of error is called a mis-stratification error. One example of how a return might be mis-stratified is the following: a corporation files a return with total assets of $10,000.23 and net income of $5,000.00; a processing error causes the cents to be keyed in as dollars so that the return is classified according to $1,000,023 total assets and $5,000 net income. The return would be mis-stratified according to the incorrect value of total assets.

The following table shows the number of mis-stratified returns for Tax Years 1991 through 1995.

Tax Year	1991	1992	1993	1994	1995
No. of Returns	1,420	1,207	1,082	1,324	1,420

Mis-stratified returns in the sample were reclassified into their proper sampling classes after complete data capture. The population of returns that needed to be reclassified was estimated from the sample and the stratum population sizes were adjusted accordingly [5]. Population and sample totals were minimally affected by reclassification, and an analysis of the sample results tended to confirm that mis-stratified returns occurred randomly. Steps are being taken by both the Centers and the SOI Division to minimize the number of mis-stratified returns.

Response errors

Response errors are due to data being captured before auditing. Some purely arithmetical errors made by the taxpayer are corrected during the data capture and cleaning processes. Because of time constraints, adjustments to a return during auditing are not incorporated into the SOI file.

Industrial Classification

The industry classification used in this report generally conforms to the former Enterprise Standard Industrial Classification (ESIC) authorized by The Office of Information and Regulatory Affairs in The Office of Management and Budget (OMB). This classification was designed to classify companies, which are often engaged in more than one industry activity, into only one industry category. The structure of this classification follows closely along the lines of the underlying Standard Industrial Classification (SIC) Manual, also authorized by OMB, which is designed as a means of classifying establishments. Some departures from the ESIC system were made by SOI for financial industries in order to reflect particular provisions of the Internal Revenue Code. The differences between the ESIC and SIC industries and the SOI industries used in this report have been analyzed.

References

[1] Jones, H. W., and McMahon, P. B. (1984), "Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present," *1984 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 437-442.

[2] Harte, J. M. (1986), "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *1986 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 603-608.

[3] Überall, B. (1995), "Imputation of Balance Sheets for the 1992 SOI Corporate Program," *1995 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 275-280.

[4] Oh, H. L. and Scheuren, F. J. (1987), "Modified Raking Ratio Estimation," *Survey Methodology*, Statistics Canada, Vol. 13, No. 2, pp. 209-219.

[5] Mulrow, J. M. and Jones, H. W. (1989), "Sampling Administrative Records: Detection and Correction of Stratification Errors," *Statistics of Income and Related Administrative Record Research: 1988-1989*, Internal Revenue, December 1990, pp. 139-144.

SOURCE: 1995, Corporation Income Tax Returns, Publication 16.