DESIGN AND USE OF AN IMBEDDED PANEL IN THE SOI CORPORATE SAMPLE

Susan Hinkins, Jeri Mulrow and Richard Collins, Internal Revenue Service
Susan Hinkins, P.O. Box 369, Bozeman, MT 59771

KEY WORDS: Cross-sectional samples, Year-to-year change estimates.

BACKGROUND

The Internal Revenue Service (IRS) has been producing statistics on economic and tax data since 1913. Up until 1950, these statistics were based on a census of all corporate tax forms that were filed with IRS. After 1950, sampling techniques were employed to produce the statistics. Since 1951, the sample size has stayed relatively constant or decreased over time, while the population size has increased dramatically (Figure 1). The overall sampling rate has decreased over the years from 40% in 1951 to just over 3% in 1987.

Some examples of the estimates produced in 1987 are:

- Total assets were equal to $15.3 trillion;
- Total liabilities, for construction companies with total assets less than $5 million, were equal to $929 million and;
- Total deductions for companies with assets of $1 under $100 thousand were equal to $305 million.

These estimates are available to the general public at the aggregate level, but are mainly used by the Office of Tax Analysis in the Department of the Treasury, the Bureau of Economic Analysis in the Department of Commerce, and the Joint Congressional Committee on Taxation. The SOI Corporate data along with data collected from other agencies are used to study the tax laws and the general well-being of the country.

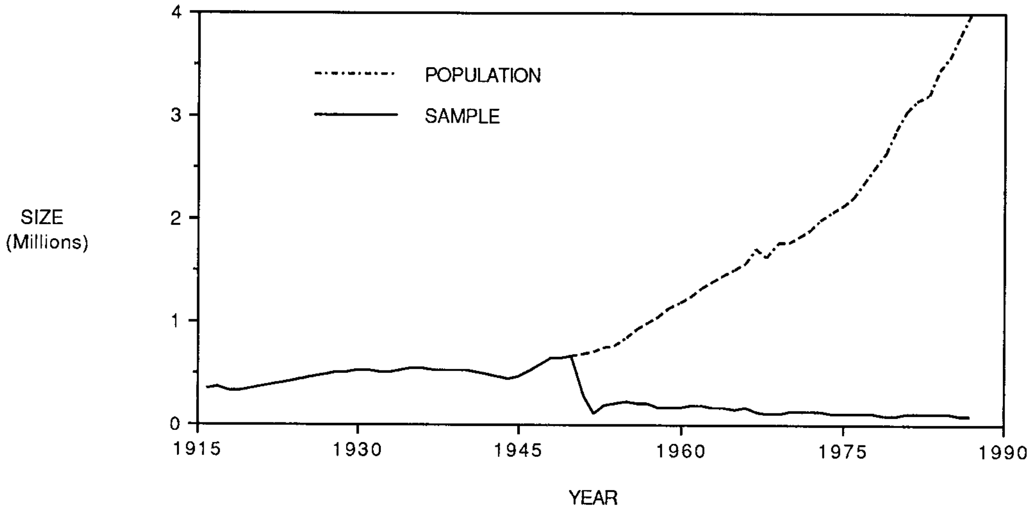
The annual or cross-sectional estimates are the primary objectives of the sample design. But there is also interest in doing longitudinal studies on corporate data and in improving estimates of change in various economic variables. To facilitate this effort, SOI began using the Taxpayer Identification Number, an individual's Social Security Number or a corporation's Employer Identification Number, as a basis for sample selection in 1968. This procedure allows for overlap of companies in year-to-year samples while retaining randomness within a given year.

From 1968 to 1978, random digits were selected in specific positions of the Employer Identification Number (EIN), a nine-digit number. Depending on the sampling fraction, three digits (positions 6,7, and 8), two digits (positions 7 and 8) or one digit (position 8) was used. The first two digits, which indicate district and the ninth digit, which was found to have a high proportion of zeros and fives, were not used for sample selection purposes.

The corporate sample was actually a cluster sample under this sampling procedure. That is, if the number 123 in digit positions 6, 7, and 8 was used as a sampling criteria, then a cluster of returns - - not a single return - - was selected for the sample. There was a concern whether the assignment of EINs to corporations caused an appreciable intra-cluster correlation to exist. Factors in the system for assigning EINs that could introduce intra-class correlation or cause some peculiarities in the sample include:

- Blocks of 20,000 consecutive numbers were reserved for agricultural and household

Figure 1. -- Population and Sample Sizes, 1916 - Present



employers;

- The EINs for some non-employing organizations called "special" are identifiable by a "6" in the third-digit position;
- Subsidiaries of a company may have consecutive identification numbers;
- The method of assignment for an EIN may have changed over the years and is not well documented before 1961, when the IRS began requiring corporations to report the number on their tax returns.

For tax year 1978, the decision was made to change the sample selection procedure to use a transformation of the EIN rather than the EIN itself. This method was first proposed and studied at the Bureau of the Census by B. J. Tepping. The general formula for computing the transform is:

$$y = c\ x\ (\mathrm{mod}\ p),$$

where

y is the transformed number and equals the remainder wher $c\ x$ is divided by p;

x is the EIN;

p is a large prime number; and

c is a constant which is relatively prime to the number of subsets the population is partitioned into.

This transformation accomplishes two important purposes: (1) the transformed number is pseudo-random, and (2) the transform, corresponding to a given EIN, is always the same. That is, the companies in the sample from a given stratum will be a random sample. If p and c remain constant over the years, then the sample has the following properties:

- The sample is self-adjusting for births and deaths.
- Tax returns for a large proportion of the corporations will be in the sample from year to year. Consider corporations selected into the first year's sample with sampling rate r, and that still exist the following year (i.e., file a tax return). Then, corporations that change to a stratum with higher sampling rate will remain in the sample the second year. For corporations that change to a stratum with lower sampling rate q, q less than r, we expect approximately (q/r)*100% to remain in the sample the second year.

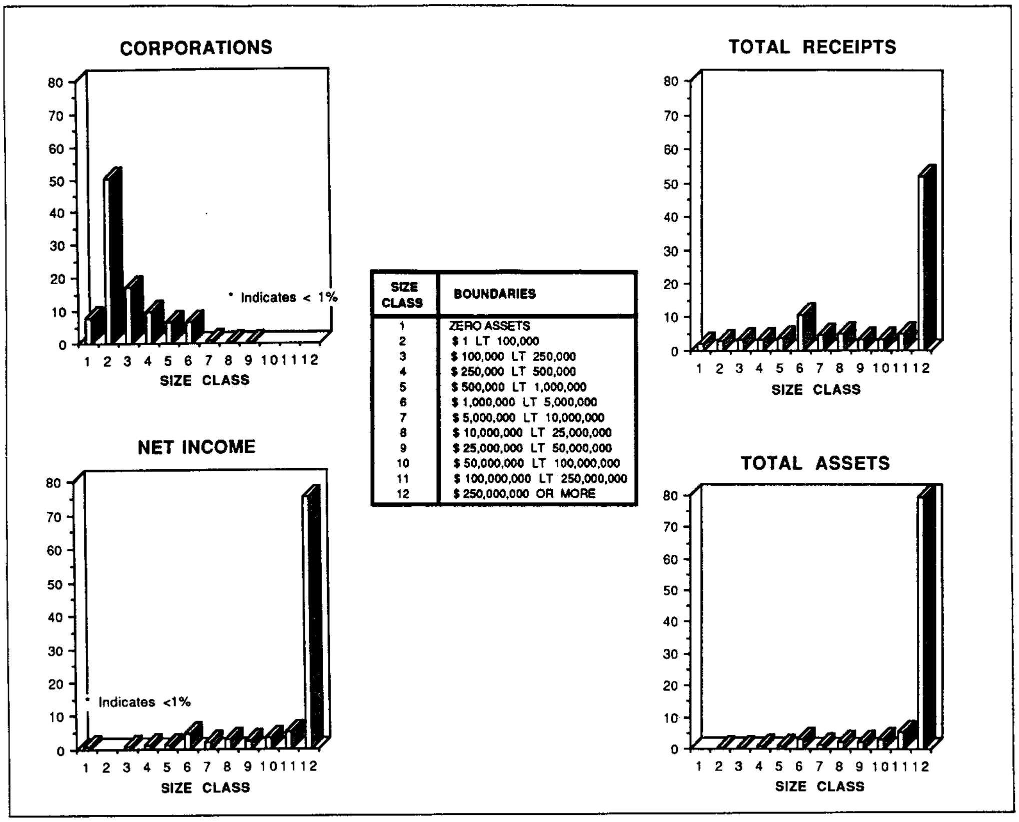
These are the same properties the sample had when specific digits in the EIN were used as the selection criteria.

In this paper we show some properties of the sample overlap from the most current six years of available data, 1982-1987. We also outline our plans for further studies.

DATA ATTRIBUTES AND SAMPLE SELECTION

As Figure 2 shows, a very small percentage of the corporate entities accounts for a large

Figure 2. -- Percent of Returns by Size Class



SIZE CLASS	BOUNDARIES
1	ZERO ASSETS
2	$ 1 LT 100,000
3	$ 100,000 LT 250,000
4	$ 250,000 LT 500,000
5	$ 500,000 LT 1,000,000
6	$ 1,000,000 LT 5,000,000
7	$ 5,000,000 LT 10,000,000
8	$ 10,000,000 LT 25,000,000
9	$ 25,000,000 LT 50,000,000
10	$ 50,000,000 LT 100,000,000
11	$ 100,000,000 LT 250,000,000
12	$ 250,000,000 OR MORE

percentage of the total dollar amounts. Over 1/2 of the population has reported total assets under $100,000. A little over 0.1% of the population holds over 79% of the total assets, 51% of the total receipts, and 75% of the income. Because the population of corporate returns is so highly skewed, the sample design stratifies on measures of the size of the corporation and selects, with certainty, large corporations.

While industrial classification is not used in sample selection, it is often of interest. Figure 3 shows the relative composition of the population and the sample by industry. The Services Industry (Advertising, Auto Repair, Doctors Offices, etc.) comprise 32% of the population but only 13% of the sample; these corporations tend to be relatively small. The Finance Industry tends to include larger corporations; it comprises only 14% of the population but 30% of the sample.

Figure 3. -- Industry Composition

	COMPOSITION	
INDUSTRY	POPULATION	SAMPLE
Agriculture, Forestry, and Fishing	3.0%	2.4%
Mining	1.0%	1.7%
Construction	9.0%	7.7%
Manufacturing	7.5%	16.0%
Transportation and Public Utilities	3.0%	4.1%
Wholesale and Retail Trade	23.0%	23.3%
Finance, Insurance, and Real Estate	14.0%	31.7%
Services	32.0%	12.9%
Not Allocable	7.5%	0.2%

Figure 4 shows the sample sizes for the years 1982 through 1987. These numbers exclude any special studies performed during those years and any duplicate EINs in the files due to the filing of both part-year and full year returns during the sampling period.

Each year the sample design included a census of all corporations having total assets (TA) of $100 million or more, and a sample of smaller corporations. The number of large corporations included in the census has grown each year. At the same time, due to cost and timeliness concerns, the total sample size has not grown but tended to decrease. The large corporations are therefore an increasingly large proportion of the total data file and the sampling rates for the smaller corporations have necessarily been decreasing. This affects not only cross-sectional estimates for the subpopulation of smaller corporations, but also the properties of records retained in the sample from year to year.

Figure 4.--Sample sizes for 1982 through 1987

Year	Total sample size	Total assets GE $100 million	
		number	percent
1982	92,924	6,783	7.3%
1983	89,594	7,257	8.1%
1984	85,518	7,611	8.9%
1985	87,742	8,686	9.9%
1986	82,658	9,423	11.4%
1987	80,266	9,712	12.1%

OVERLAP OF SAMPLED CORPORATIONS

The effectiveness of using the EIN in retaining a large sample overlap from year to year is determined by 1) dynamics of the corporate population, and 2) dynamics of the sample design and operation. Corporations change over time and therefore may move to different sampling strata. They may grow or decline in prosperity, change the nature of their business, merge, go out of business, etc. There may also be changes in the corporate record (tax return) due to tax law changes or new tax forms. Corporations may be non-filers (and, therefore, out of our population) because they merged, or did not need to file one year, or permanently went out of business.

The properties of the sample overlap also depend on dynamics of the sample design. Changing the definition of the sample strata and changing the sampling rates will obviously affect the sample overlap. Operational difficulties that affect the sample selection also affect the sample overlap. Most noticeably, data errors in the stratifying variables at the time of selection result in mis-stratification of records and, therefore, in selection errors. (Mulrow and Woodburn, 1990.)

In this section we will look at a few properties of the sample overlap for 1982-1987. Over this time, all these possible population and sample design factors were affecting the sample overlap. We consider the result; in subsequent work we will try to isolate the effect of particular factors on the overlap.

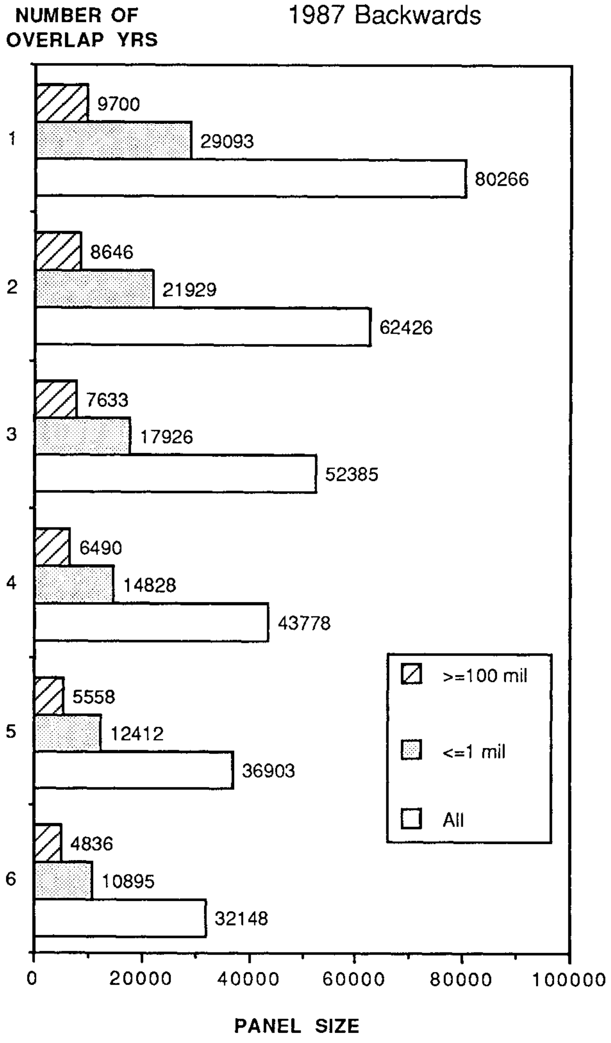
Figure 5 gives a retrospective look at the imbedded panel of corporations from 1987 back to 1982. That is, suppose we start with the file for the 1987 sample and want to see for which records we have historical (panel) data. As expected, the longer the time period desired, the smaller the resulting panel:

- 78% of the corporations in the 1987 database were also in the 1986 database;
- 65% of the 1987 corporations were in both 1986 and 1985 databases; and
- 40% of the 1987 corporations were in all 6 years, back to 1982.

Since large corporations are selected with higher sampling rates than small, the use of the transformed EIN for sample selection makes it more likely to keep growing corporations in the sample over time, than corporations that decline in size. Therefore we looked at three groups of records in the 1987 file:

- large corporations, those with 1987 total assets (TA) $100 million or more (note that these were selected with certainty in the 1987 sample);
- medium size corporations, those with 1987 TA between $1 million and $100 million; and
- small corporations, those with 1987 TA less than or equal to $1 million.

Figure 5. -- Retrospective Overlap
1987 Backwards



As expected, the largest 1987 corporations have the relatively largest panel:

- 50% of the largest 1987 corporations have data for all 6 years;
- 42% of the middle-sized 1987 corporations; and
- 37% of the smallest corporations are in the 6 panels.

PROSPECTIVE VIEW OF THE PANEL DATA

We considered the panel data over a shorter time interval, 1984-1987, and starting with the 1984 sample we looked at how these records changed over time. Again, categories of corporations were defined by size of total assets: 1) the smallest corporations in 1984, and 2) the largest 1984 corporations. The panel data for small corporations consist of 13,560 corporations selected in the 1984 sample with TA between $100 and $1,000,000 that were also in the 1985-1987 samples. Similarly the panel data for large corporations consist of 5,292 records that had 1984 TA over $100 million and were selected in the 1984-1987 samples.

We look at how TA changed over time. Figure 6 shows the percent change in TA for each data set, first after one year and then after 3 years. The percent increase, for, say, the 1984 to 1985 change, is calculated as

100 * [(1985 TA) - (1984 TA)] / (1984 TA).

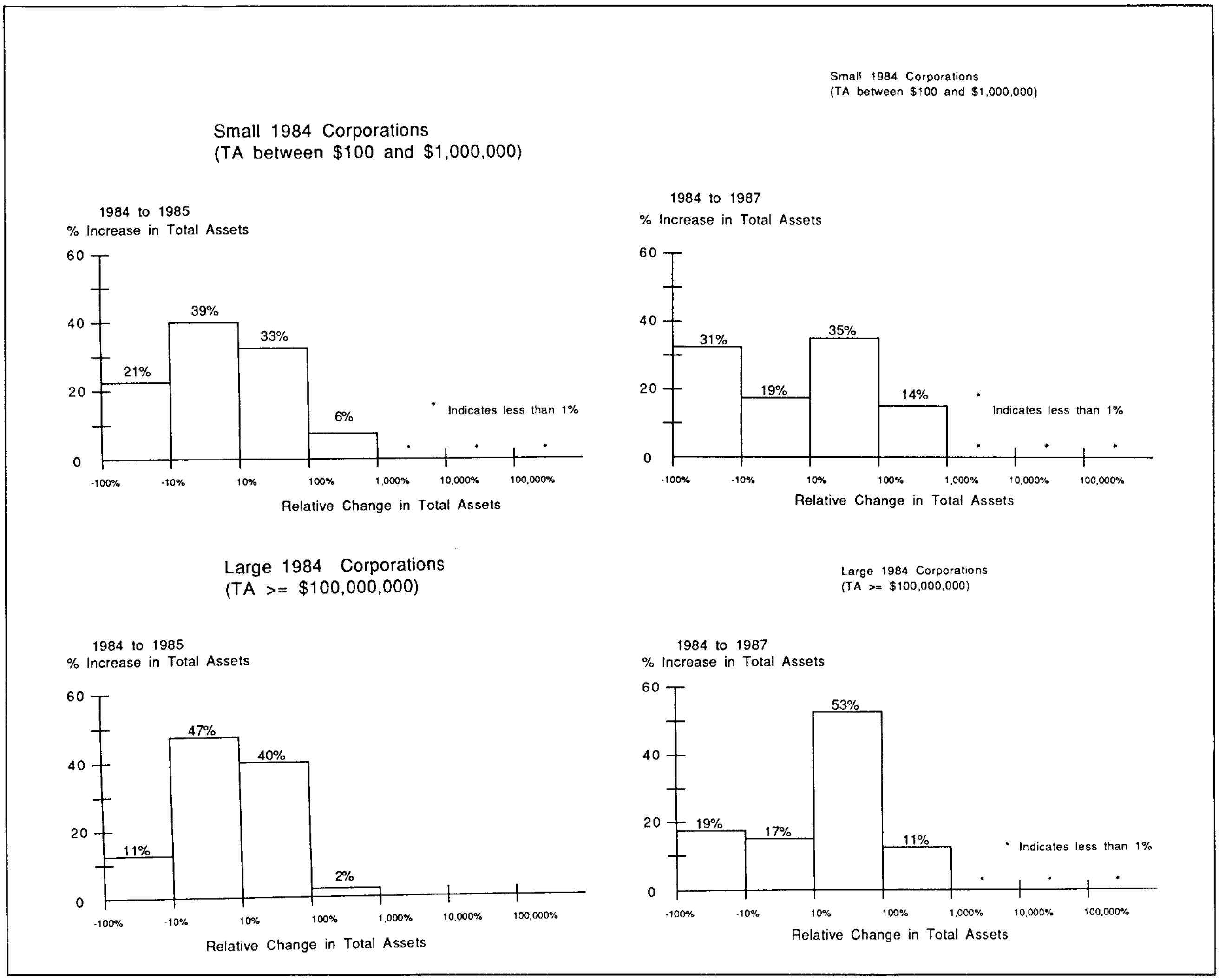
Therefore, for companies with decreasing TA, the percentage "increase" is bounded below by -100% since TA must be nonnegative. However, there is no such upper bound for corporations with an increase in TA.

It is important to remember that the following description and summary are for the panel data, which are not a random sample either of the cross-sectional samples or of the population.

Consider first the change after only one year. For both the smallest and largest corporations the panel data contain a large percent of records with essentially no change in TA: 39% and 47%. Both groups show about the same percent of corporations that grew (i.e., a relative increase over 10%), though the small corporations show more very large relative changes. The major difference is that the small corporations show a greater tendency for decreases in TA than the large corporations. This tendency is more noticeable when considering the change from 1984 to 1987.

In each category of corporations, there were more corporations with changes after 3 years than after one year. Comparing changes in the

Figure 6. -- Relative Change in Total Assets



small corporation panel after 3 years to the large corporation panel:

- Small corporations were more likely to have a decrease in total assets of -100% to -10% than the large corporations (31% vs 19%);
- Small corporations were less likely to have an increase of 10% to 100% than the large corporations (35% vs 53%); and
- The percentage with no change in TA and the percentage with larger changes were essentially the same in both groups.

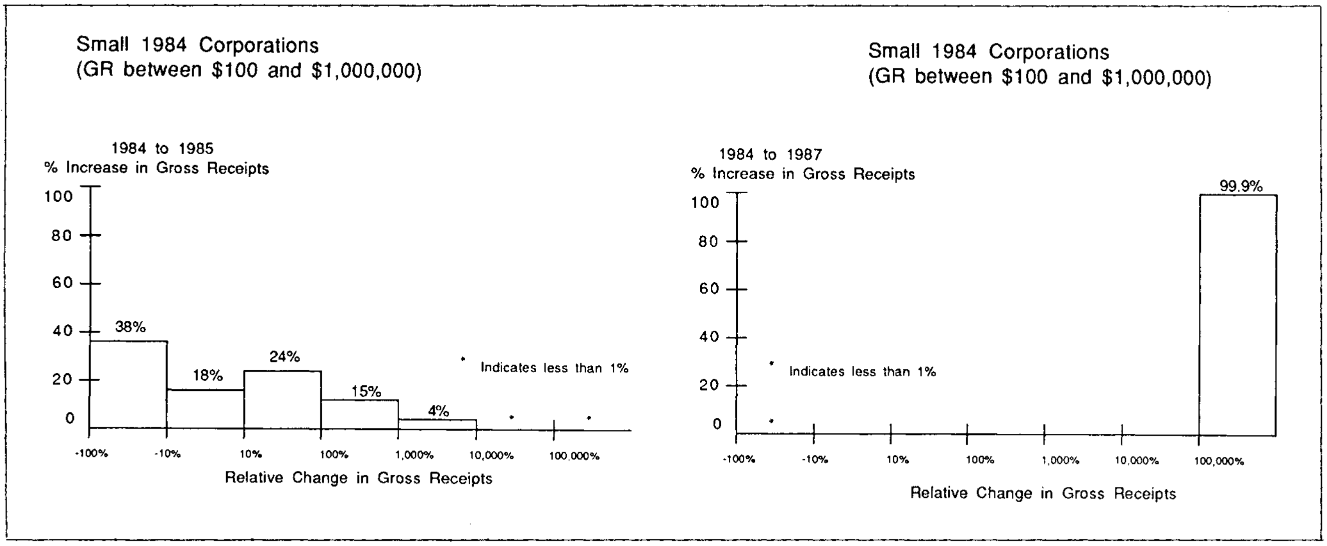
We know the panel data are a non-random sample. Looking at TA, a stratifying variable, we know something about the process. Holding other factors constant, corporations that steadily increase in TA are more likely to remain in the sample than corporations that are static, and corporations with relatively static TA are more likely to remain in the panel than those with a decrease in TA in even one year.

Figure 7 considers the change over time in a different subpopulation of the panel: corporations that in 1984 had gross receipts between $100 and $1,000,000. A rather surprising (relative) change in gross receipts is seen over time. According to subject matter experts, this substantial increase from 1984 to 1987 is believable; these were originally small values of gross receipts in corporations that then survived for 4 years. However, we do not know to what extent the nonrandom aspects of the panel data may have influenced such an extreme result. By weighting the panel data correctly, we can estimate population distributions. But when the panel data give an extreme distribution, as for gross receipts, with so many empty cells, then it is difficult to know to what extent this is a property of the nonrandom selection of the sample overlap.

FUTURE PLANS

We are just beginning 1) to evaluate the current method of assuring a large sample overlap at little or no cost to the cross-sectional estimation and 2) to investigate the characteristics of the resulting sample overlap. Therefore, our future plans still

Figure 7. -- Relative Change in Gross Receipts



comprise the majority of this work.

We plan to evaluate the use of the transformed EIN in sample selection by comparing properties of corporations selected using the transformed EIN with properties of the original population. For example, we know that for small sampling rates, if a given EIN, say N, is selected into the sample, then the probability of selecting the record with EIN=N+1 into the sample is zero (Harte, 1986.) The first two digits of the EIN represent the district that assigned that corporation its EIN. We know that the distribution of corporations over districts varies with different properties of the corporations: industrial classification, size, etc. Therefore, in a simulation study, we will compare the distribution of districts in sampled records to the population distribution. We also hope to compare the earlier methods of sample selection, based on digits of the EIN, to the method using the transformed EIN.

As mentioned previously, we want to try to estimate the effects of certain isolated factors on the loss of corporations from the sample overlap. For example, if we kept the strata definitions constant, how much of the sample would be lost due to corporations not filing, how much due to corporations changing strata, and how much due to sample rate reductions?

Finally, we need to evaluate the characteristics of the panel vs the cross-sectional samples and vs the population. And we need to evaluate the benefits and costs of certain design changes, recently made or considered, in terms of the effects on the year-to-year overlap, the cross-sectional estimation, and the operational costs.

ACKNOWLEDGMENTS

The authors wigh to thank Paul McMahon and John Epperson for supplying population data, and Beth Kilss for her help in developing the ASA posters and this paper.

REFERENCES

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

Hinkins, S. and Scheuren, F. (1989). Evaluating Sample Design Modifications: Balancing Multiple Objectives, Proceedings of the American Statistical Association, Business and Economic Statistics Section, 654-658.

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

Mulrow, J. (1990). Description of the Sample and Limitations of the Data, Statistics of Income... 1987, Corporation Income Tax Returns, Publ. 16, 9-15.

Mulrow, J. and Woodburn, L. (1990). An Investigation of Stratification Errors, Proceedings of the American Statistical Association,Section on Survey Research Methods.

Tepping, B. J. (1969). Memo to Mr. J. F. Daly, dated January 15, 1969, U.S. Departement of Commerce, Bureau of the Census.