Sample Design Revisions in the Wake of NAICS and Regulatory Changes

Paul B. McMahon, Internal Revenue Service, P.O. Box 2608, Washington, DC 20013
paul.b.mcmahon@irs.gov

Key Words: Administrative Records, Sample Design, NAICS

Introduction

Many establishment surveys depend upon administrative record sets as the basis of the sampling frame. A favorite of frame builders is a file of tax records, when it can be obtained, because the records seem rich in possible stratifying variables. These data sets, however, are subject to changes depending on the regulations that gave rise to their existence. Often these changes, like in the industry classification, are known in advance at least in outline. Another factor, though, is the hidden implicit strata breaks that some of these sets have that might not be at all evident from the documentation. One example of this is in the data abstraction procedures for corporation tax returns, which handles firms with more than $10 million in assets differently than those with less. Since strata are devised for homogeneity, this difference in treatment implies the need for strata that respect this boundary.

We examine the Statistics of Income Partnership Study for examples of these situations and their effects. This study has been conducted annually for about 50 years, with varying sample designs, relying on the data in the Internal Revenue Service's Business Master File System for classification information. For the past 25 years, the designs have used the industry code as a primary stratifier, along with asset size and receipts. The change from the Standard Industrial Classification (SIC) based industry codes used by the IRS to a set based on the North American Industry Classification System (NAICS) threw a well-established outline into some disarray, while the recent reorganization of the Service has had its own effect.

But first, to set the stage, we will begin with a brief description of the population of interest and the administrative environment in which this design must operate. We will then review the impact of the switch in industry coding on the existing design, and close with an outline of the modifications we are putting in place.

Background

The establishments we are interested in are businesses that have more than one owner, are not incorporated, and are required to file a Partnership Return of Income, Form 1065, with the Internal Revenue Service annually. This population does not include operations conducted under joint operating agreements, as are sometimes used by lawyers sharing office space or oil companies sharing a drilling rig. Yet it does include things like Limited Liability Companies and Publicly Traded Partnerships that, to a non-lawyer, would certainly appear to be corporations.

The Office of Tax Analysis, of the Department of the Treasury, and Congress's Joint Committee on Taxation, the sponsors of this study, are primarily interested in reviewing tax laws, but these businesses, the partnerships, are not usually directly taxed. The reason for the attention is that these entities are conduits for profits, expenses, and various tax credits to be allocated to the owners. This allocation of credits, and so on, is determined by the partners, not by regulations, and need not be equally shared. This form of business organization, as a result, is often used in the creation of tax shelters, which of course draws the eyes of our sponsors.

The source of the data for these studies is the Partnership Return of Income, Form 1065, filed by each firm in the population. Selected information on that administrative record is transcribed onto electronic media (or edited from one version to another), then posted to the Business Master File.

There are four types of fields present on the Internal Revenue Service's Business Master File System, which serves as our sampling frame: administrative, entity, codes, and amounts. The administrative fields contain only items like the work group and audit trail data, which are of no interest to our clients.

The entity data include items like name and address and have a few items, particularly the State in which the firm was organized, that can be of occasional use. However, the Statistics of Income Corporation and Partnership studies are designed for national estimates, so selected State estimates are only rarely produced.

The code fields are answers to questions about foreign owners, nature of the accounting methods, nature of the organization (such as whether it is a limited partnership), and other categorical information. The NAICS code is among these--or rather, the IRS's version of them. The list used is, for the most part, a partial collapsing of the 1,170 NAICS classes for United States businesses, resulting in about 420 codes. The number of codes depends upon the type of organization, for various laws prohibit certain businesses from incorporating (accounting firms, for example), while requiring it of others (insurance companies). (The list for corporations and partnerships may be found in the instructions for the forms.)

From a cross referencing perspective, of more particular interest is that the sampling frame has, for the past 3 years, contained information about the industry code used on previous filings—the last SIC

based code reported. This is not a validated code, and, like the IRS's NAICS Codes, it was selectively edited from the full list of SIC Codes. This information will be preserved on the population files maintained by the Statistics of Income Division, though removed from the Business Master File after December 2001.

There are a relatively small number of amount fields, compared either to the number needed by our sponsors (we collect about 300 items for them) or to the potential number on the form and all the various attachments. Depending on the year of the record's creation, there are about 40 monetary variables present for possible use in stratification.

From the design standpoint, we need fields that are highly correlated to the data of interest, but not to each other. Many of the fields in the records are very highly correlated. For example, cost of goods sold and net receipts have a correlation coefficient that is very close to 1 (about 0.99). This is only to be expected, given the structure of the accounting data we are dealing with. That structure also has an industry component to it, for income from real estate rent is not part of net receipts, and related deductions are not included in the calculation of net income.

This arises out of the division of sources of income into "active" and "passive," which is a legacy of the 1986 Tax Reform Act. Rents, like income from a portfolio of stocks and bonds, are considered "passive." This distinction is included in the tax law as a way to discourage the formation of tax shelters.

This dividing of income sources, though, also had the effect of creating some income fields that are essentially a proxy for a firm's industry. Real estate rent is one example. For our needs, then, we must have a consistent economic (rather than tax law) definition of either net income or total receipts. As a result, several fields are combined for stratification purposes, coming as close as we can to those economic measures.

Tax Year 1997 Sample Design

Since the Partnership study is conducted annually, we prefer to use, as nearly as possible, the same outline from one year to the next. This minimizes complications that arise in analyzing the changes between years and, incidentally, makes the maintenance of the computer operations simpler.

Those computer operations, present a planning challenge, for we must integrate the sample selection procedures into the IRS's processing. This puts our planning requirements on their schedule, which is important to this story.

In February 1998, nearly a year before the first Tax Year 1998 return was due to be filed with the new NAICS industry information, we had to finalize the sample design for that year. In January 1999, the first returns were filed and subjected to sampling. The selection continued throughout 1999, but the data abstraction and editing for the 1998 Study were not completed until April 2000. That is, the first data on the NAICS distribution became available 2 months after we were committed to the design for the Tax Year 2000 Study.

In the design for Tax Year 1998 (selected during 1999), we had little but the descriptions of the new industry codes to go by. Thus, we first look at an outline of the Tax Year 1997 design as the pattern for the studies, then at the translation used to make the interim modifications for the first NAICS selections.

Figure 1: Pre-NAICS Design for the Statistics of Income Partnerships Studies

Very Large Cases
(2 Strata)

Real Estate Operators (Except Developers) and Lessors of Buildings
(SIC 6511, 18 Strata)

Agriculture, Trade, Finance and Services
(SIC's not specified elsewhere, 29 strata)

Mining, Construction, Manufacturing and Transportation
(SIC's 1000 through 4999, 19 Strata)

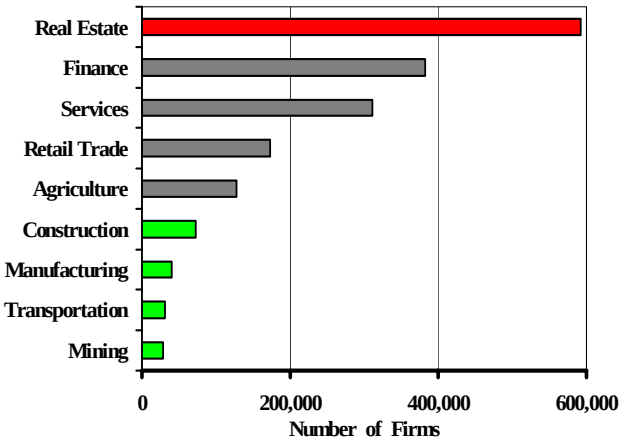
The Very Large Cases were those with either $100 million or more in total assets, or $25 million or more in either the computed receipts or income measures. All of the strata within the industry groupings were classified on these same characteristics, and given the structure of the data on the sampling frame, there is no reason to modify this approach.

The change in the industry classification system, though, does invite investigation into the rationale behind the choice of industry groupings.

The data in Figure 2 are estimates from the Tax Year 1997 study. The graph clearly shows that the single industry, Real Estate Operators (except Developers) and Lessors of Buildings (in red), dominates the Partnership population with about a third of the firms. If a proportional sample allocation were used, then about 12,000 records would be used to provide less than two percent of the total number of published estimates. At the same time, 3,400 (total) records would be used in the estimates of more than 20 percent of the estimates for the, Construction, Manufacturing, Transportation and Mining industry divisions (green, in the figure below).

Since our sponsors are interested in various industries at different times, we need better distributional properties across industries than this. At the same time, we need to retain decent income and

Figure 2: Tax Year 1997 Partnerships: Industry Distribution



asset distributions. Our solution, first introduced for the Tax Year 1977 study [1], was to separate the Real Estate Operators into their own strata and restrict the sample allocation to about half the proportionate share. We also provided more strata and about doubled the sample size for the smaller industry divisions, beginning with the Tax Year 1993 study [2].

Tax Year 1998 Design Modifications

When the planning for the Tax Year 1998 Study began, in late 1997, there were no data on what the migration from the SIC-based industry coding to the NAICS-based coding would yield with respect to the industries distribution. A good proportion of the firms did not even exist at that time, the filing period was more than a year off, and the tax forms themselves had not been created.

Lacking any information, then, we assumed that the same distribution would be present and tried to use the NAICS descriptions for a conversion. This conversion is shown in Figure 3, below. This strata plan, with the associated sampling rates (used in the Bernoulli selection procedure [3]), was transmitted to the programmers in February 1998.

There were a couple of other changes to the design. First, the number of largest firms had grown to the point that we decided to raise the boundaries of the certainty classes to $250 million in assets (up from $100 million), and to $50 million for net income or receipts (up from $25 million). We installed two new strata to fill the gap with the blocks of industry classes, and sampled them at a 50-percent rate. The other two modifications arose from a regulatory change.

A new form was introduced, the 1065-B, that was to be used by companies with 100 or more partners. Unfortunately, the rule for abstracting amounts from this new form was quite abbreviated: no money amounts other than remittance (in the rare case that any money was due).

The second administrative change was even smaller. In order to identify Publicly Traded Partnerships, a special value was inserted in one of the existing audit trail codes. Our sponsors were eager to review these firms' reports, and since they were thought to number only a sparse handful, we took advantage of this opportunity. We only learned of this coding plan late in the process, far too late to provide another stratum for these firms.

Since we believed that there would not be very many filers, and that they would likely have been among the largest firms as well, we created a separate class and selected all of them for the sample.

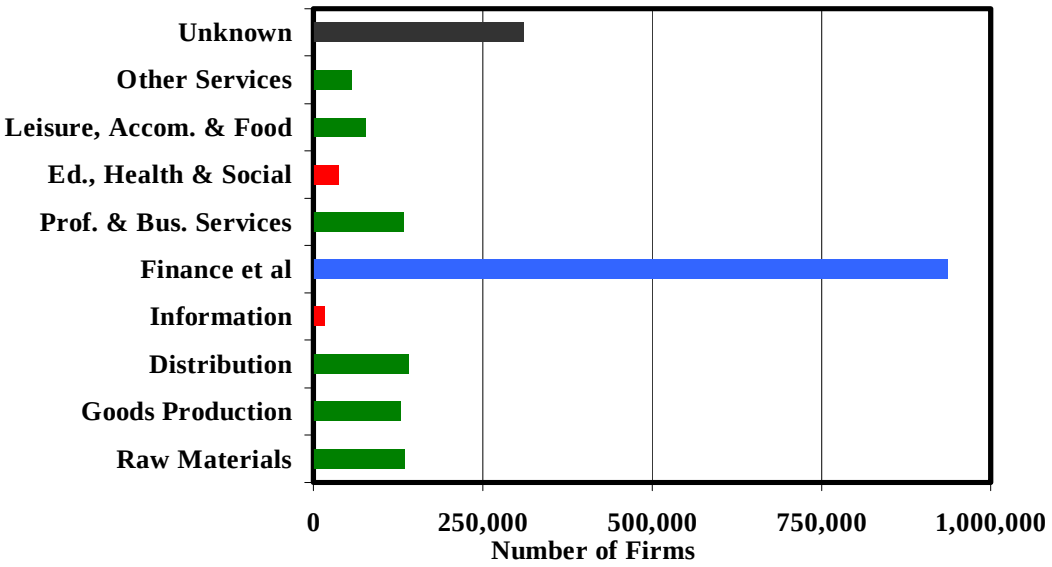
At the end of Calendar Year 1999, we saw how this played out. We were slightly over our target of 35,000 active firms, but this was due, in part, to clear coding errors. About a thousand records were processed as the new form for firms with very large numbers of partners, but very few really were that large.

A larger problem was that the revised industry groups did not fit the new industry classification.

Figure 3: Industry Groups Used in the Tax Year 1998 Sample Design

	Principal Business Activity Codes	
	Standard Industrial	**North American Industry**
Industry/Division	**Classification**	**Classification System**
Real Estate Operators	6511	531110 and 531120
Mining, Construction, Manufacturing, and Transportation	1000 through 4999	200000 through 350000, and 480000 through 519999
Farms, Trades, Finance, and Services	All Other Codes	

Figure 4: Partnerships by NAICS Industry Divisions, Tax Year 1998



NAICS Industry Distribution

The population data that gave rise to the NAICS Industry chart, above, became available in mid-January 2000, too late to affect a design before the Tax Year 2001 cycle (with the sample to be selected during 2002).

There are three items in Figure 4 that are key: the size of the "Unknown," the dominance of the Finance division, and the presence of a few small divisions. The Unknown arise from the returns that are filed for previous tax years, taxpayers' habit of using prior-year filings as the basis for the next, and IRS's input errors. These errors in the initial year of using NAICS codes were higher than in later years, due to some confusion by the initial input clerks.

How reliable are the NAICS codes on the Internal Revenue Service's files? The data in Figure 5 are from the raw input files, which may cause some records to be counted more than once. We do not, as of this writing, have complete data for Tax Year 2000 for the simple reason that most of the records have not been received and processed yet. Those figures are for records processed through the end of August 2001.

Figure 5: Percent of NAICS Codes Validity on Partnerships Returns, by Tax Year

Tax Year	Valid	Not Supplied	Invalid	SIC
1998	80.8	5.0	8.2	6.1
1999	85.0	4.6	7.4	3.0
2000	86.2	4.3	7.4	2.1

"Valid" is defined here as being one of the industry codes that IRS includes in the instructions associated with the Partnership reporting form. This is not the complete list of codes, but a reduced set combining many rare industries. A fair proportion of the "Invalid" codes cited above are likely to be acceptable codes to other agencies, but the source of the above data does not contain this information. The data on the "Not Supplied" are very close to the number of firms that show no current activity, which was about 4.5 percent for the Tax Year 1999 Study. Still, the data for the early Tax Year 2000 returns are encouraging, as later filings tend to have a somewhat higher proportion of valid codes than those filed earlier.

The former sparsely populated SIC divisions of Mining, et al. are no longer present under NAICS. Indeed, the choices for the replacement industries led, in part, to missing our target sample size by about 20 percent. The new distribution has its own small divisions, however--Information, and Education, Health and Social Services (Figure 4, in red). (We considered "Other Services," too, but there was not sufficient interest from the subject-matter specialists, to warrant an elevated sample size.) These industries replace the SIC-based sparse divisions in the Tax Year 2001 design.

At the other end of the spectrum are the highly populated single industries. If any are to be found, they are in the Finance Division, and, in fact, there are four candidates.

Figure 6: Largest Finance Division Industries

	Firms
Other Financial Investment Activities	113,500
Residential Buildings and Dwellings	285,300
Non-Residential Buildings	237,000
Other Activities Related to Real Estate	116,700
All Other Finance Industries, Total	184,100

Real estate businesses, under NAICS, are no longer confined to a single industry, and are now about 3 percent of the published estimates. However, the tax attributes that made them the dominant group had not simply vanished; they all still file the same attachments, particularly Form 8825, Rental Real Estate Income and Expenses of a Partnership or an S Corporation, and take the same deductions, like the depreciation on buildings. Thus, the rationale for separate real estate strata is still sound, as is the reduction in sample resources from a proportional allocation.

In the case at hand, we selected the inheritors of the old SIC industry, as determined by a review of the migration [4]. These were: Lessors of Residential Buildings and Dwellings (531110), Lessors of Nonresidential Buildings (except Miniwarehouses) (531120), and Other Activities Related to Real Estate (531390).

Monetary Strata

Within the three industry classes, the records on the sampling frame are categorized by size of total assets, and the larger of receipts or absolute value of net income (loss), as shown in Figure 6 (along with the sampling rates for Tax Year 2001). The boundaries for the classes were not entirely of our choosing, as, once again, regulations come to the fore.

Since strata are designed to be as homogeneous as possible, if a regulation treats some members of a population differently, then that regulation is effectively setting strata boundaries. There are three that appear in the design revision: two arise from an exemption on reporting details of asset holdings, another from organizational alignment.

On page 2 of the 1998 version (Schedule B) of this form is the question below:

*"**5** Does this partnership meet **ALL THREE** of the following requirements?*

***a** The partnership's total receipts for the tax year were less than $250,000;*

b** The partnership's total assets at the end of the tax year were less than $600,000; **AND

***c** Schedules K-1 are filed with the return and furnished to the partners on or before the due date (including extensions) for the partnership return.*

If 'Yes,' the partnership is not required to complete Schedules L, M-1, and M-2; Item F on the front page of Form 1065; or Item J on Schedule K-1."

The boundaries for total assets and receipts, shown in Figure 7, reflect this reporting exemption. This exemption affects 47 of the key data elements we abstract. In effect, this is regulation generated item nonresponse, and since whole schedules are affected, a weighting scheme can be effective. This in turn suggests certain efficiencies if the adjustment cells coincide with strata.

The other boundary is not apparent from reading the filing instructions or forms, but arises out of the IRS restructuring around operating divisions that concentrate on different types of taxpayers. One of these new divisions is "Large and Mid-Size Businesses," which had plans to process firms with Total Assets of $5 million or more at a single site under their organization.

At the moment, there are no plans to process these firms' reports differently than smaller companies' filings. However, we plan to retain this design structure for several years and the process may change.

Indeed, well after the design was finalized there has been such a change. As of October 1, 2001, the boundary for the Large and Mid-Size Businesses was raised to $10 million. Unfortunately, we cannot amend the design at this late date, so this constraint will have to wait for inclusion in a few years.

Notes and References

[1] The description of the sample for the Tax Year 1976 study had 6 strata, without industry classification shown. The following year, there were 14 strata, and 7 required the presence of an SIC Code of 6511. See Internal Revenue Service, Statistics of Income–1976 Business Income Tax Returns, page 427, U.S. Government Printing Office, Washington DC, 1979, and Internal Revenue Service, Statistics of Income–1977 Partnership Returns, page 3, U.S. Government Printing Office, Washington DC, 1980.

[2] McMahon, Paul (1995), "Statistics of Income Partnership Studies: Evaluation of the Expanded Sampling Plan," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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

[4] McMahon, Paul, (2000), "Changing Industry Code Systems: The Impact on the Statistics of Income Partnership Studies," *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association.

Figure 7: Tax Year 2001 Partnership Sample Design and Sampling Rates

Extreme and Special Cases:

Total Assets $250,000,000 or more, or Receipts or Net Income $50,000,000 or more 100%

Publicly Traded Partnerships or Firms With 100 or more Partners 100%

Total Assets 100,000,000 Under 250,000,000 and Receipts or Net Income Under 50,000,000, or
Total Assets Under 100,000,000 and Receipts or Net Income 25,000,000 Under 50,000,000 . . . 35%

Real Estate

Assets ($)	Absolute Value of Receipts/Income ($): Under 50,000	50,000 under 100,000	100,000 under 250,000	250,000 under 500,000	500,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 under 25,000,000
Under 250,000	0.12%	0.20%	0.30%	{ ←	1.50%	→ }	↑
250,000 under 600,000	0.17	0.19	0.30	{ ←	1.10	→ }	
600,000 under 2,500,000	{ ← 0.27 → }		0.35	0.50	{ ← 1.50 → }		10%
2,500,000 under 5,000,000	{ ←	0.50	→ }	0.80	0.90	1.90	↓
5,000,000 under 25,000,000	{ ←	1.00	→ }	1.00	1.70	2.50	
25,000,000 under 100,000,000	{ ←			7.0%		→ }	15%

All Other Industries

Assets ($)	Under 40,000	40,000 under 100,000	100,000 under 250,000	250,000 under 1,000,000	1,000,000 under 2,500,000	2,500,000 under 5,000,000	5,000,00 under 25,000,000
Under 200,000	0.35%	0.50%	0.75%	0.12%	{ ←	3.8% → }	↑
200,000 under 600,000	0.40	0.80	0.95	1.40	{ ←	2.50 → }	
600,000 under 2,000,000	{ ← 0.65 → }		0.95	1.80	3.00	4.50	14.%
2,000,000 under 5,000,000	{ ← 1.50 → }		2.50	3.00	{ ←	6.00 → }	
5,000,000 under 10,000,000	{ ←	2.50	→ }	3.00	5.00	6.50	↓
10,000,000 under 25,000,000	{ ←	5.00	→ }	{ ←	6.00	→ } 10.00	
25,000,000 under 100,000,000	{ ←			14.%		→ }	30.%

Information, and Health, Education and Social Services

Assets ($)	Under 40,000	40,000 under 100,000	100,000 under 250,000	250,000 under 500,000	500,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 under 25,000,000
Under 150,000	0.35%	0.90%	1.50%	1.50%	{ ←	3.50% → }	↑
150,000 under 600,000	{ ← 3.00 → }		20.0	{ ←	3.00 → }	4.00	
600,000 under 5,000,000	{ ← 4.00 → }		12.0	{ ←	3.00 → }	7.00	13.%
5,000,000 under 25,000,000	{ ←	25.0	→ }	{ ←	20.0 → }	7.00	↓
25,000,000 under 100,000,000	{ ←			40.%		→ }	30.%