Administrative Records, Regulations, and Surveys

Paul B. McMahon, Internal Revenue Service

Administrative records are of considerable use in the operation of surveys due to the presence of ancillary information that permits the development of more efficient sample designs. Yet the literature on their use often contains warnings. Almost 20 years ago, for example, the Federal Committee on Statistical Methodology (FCSM) published Statistical Policy Working Paper 6, "Report on Statistical Uses of Administrative Records." In this report's *Findings and Recommendations*, the committee warned: "From a statistical point of view, the standards of quality and consistency in administrative data collection and processing programs are frequently inadequate." [FCSM 1980, page 1]. These sentiments are echoed in the recommendations quoted below:

"Recommendation 2.—The quality of administrative records to be used for statistical purposes should be evaluated systematically to determine the appropriateness of the records for the proposed use."

"Recommendation 3.—Consistent procedures should be used in administrative and statistical data collection efforts for defining reporting units, identifying and coding reporting unit characteristics, and developing standards for data tabulation." [FCSM 1980, page 2].

The term "administrative data" is often used to disguise the fact that we are, in fact, frequently talking about tax records. Over the past couple of decades, there has been an increasing interest at the Internal Revenue Service to improve the quality of these records from a law enforcement point of view. Yet the Federal Committee's concerns are as relevant today as when written. This is due in no small part to the laws and regulations and processing rules that give rise to tax records in the first instance.

Before launching into the details of some of the process that yields these somewhat maligned records, a brief review of their characteristics is in order. How do these records come to be? We will then focus on the different types of tax records and the multiple layers of regulatory instructions that impact the utility of these files from the statistician's point of view.

■ Background

Although administrative records come in many guises, we must recognize that they are byproducts of some process; they are, essentially, audit trails. When we make a purchase at a grocery store, for example, a transaction is recorded, these days with all sorts of ancillary data on exactly what was bought. These records are combined in various ways to facilitate inventory control or measure the level of business activity from one day to the next. The records are combined over time to create monthly reports, then annual summaries, in ways that preserve some characteristics but drop others.

Similar records and processes are used to order or pay for supplies and labor, or in the payment of taxes, as well as in the acquisition or sale of assets. This recording of business activity might be used in tracking products of wildly different natures, like computers, financial instruments, or a service (transportation, health, or repairs). They are also used to process business claims (insurance of various kinds, credit accounts, or warrantees). The records are accumulated for the business's internal use, but, as we noted, they are subject to editing as well as combining. The rules that govern this reduction of the data arise out of the business's own needs, out of contractual requirements, and from various governmental rules.

One of the most prominent of these sets of rules is the tax code. But that body of law does not operate in a vacuum, for other governmental bodies have their effect on what tax records actually show.

There are five sources that affect tax records. Federal tax laws are one obvious source, but there are also

treaties with foreign governments and laws that focus mainly on non-treasury agencies that occasionally impact the reported information. Examples of this would be a transportation bill that might include a new tax on aircraft or airline tickets, or a clean air bill that is funded through an excise tax on chemicals, or when such a tax expires. These non-tax laws directly levy taxes and have a serious effect on Form 720, *Excise Tax Return.*

Then, there are the laws of the various States. It is not often understood that the rules of one State can affect a national records system, but the very existence of corporations lies within the purview of the States, not the Federal Government. Another example is the Publicly Traded Partnerships, which have interests sold in the financial markets and have the limited liability of common stock, but still retain the classification of an unincorporated business. States can, and do, require certain businesses to have particular fiscal periods. Moreover, they can force companies to close their doors until some specific conditions are met, as in the case of the Rhode Island Savings and Loan crisis of a few years ago.

The common law is also a source, like the "reasonable man rule." Here, any requirement for reporting might be said to have been fulfilled, even though the exact definitions on the administrative form were not directly met.

Then, there is the matter of the courts, which might uphold the common law interpretation as well as view legislation in an unexpected context. A judge might, for example, rule that since some point of law is "sub jure," tax forms due the IRS must await that court's decision. Sometimes, this causes records to be filed years later than they would ordinarily have been expected. (Often, you hear these called "delinquent filers," but since they are adhering to the laws, this is not accurate.)

And lastly, there are the regulations themselves. There are two levels of tax agency regulations that affect surveys. The first is that set of rules describing what "respondents" must do. These define terms, set deadlines, impose penalties, and otherwise set forth what is needed for compliance with the laws. This is what makes the records valuable, but the massive volumes are hardly fun reading. They also serve as the starting point for the creation of those "questionnaires" we also know as Tax Forms.

The other regulations are internal, affecting the ways in which the Service interacts with the public, for example. More importantly for our purposes, though, these rules describe what elements from the tax form questionnaires are included on the computer files. This varies, depending on the type of record we are interested in for a particular study.

■ Uses in Surveys

Some people, hearing that tax records are provided to other agencies, conclude that the reverse is also true. It is not, but non-tax agencies are wise to avoid stating their use of tax records bluntly, even when those records are in the public domain (as in the case of Tax-Exempt Organizations). But what do they use these records for?

Because the data are reasonably well-defined, response rates are high (it ***is*** a voluntary system, after all), and the data are subject to verification, we assume these to be factual reports. There are four main uses: verification of survey responses, benchmarking, as the source of a sampling frame, and as the survey instrument.

Verification studies rely on comparisons between respondent data and those from the administrative record on an item-by-item basis. In this, we also include the replacement of missing respondent data with administrative proxies. The Internal Revenue Service's Statistics of Income (SOI) Division, however, does not conduct verification studies as a rule. Since others have covered this topic in considerable detail, and we have little experience in this matter, we will not comment further on this use.

Benchmarking has two roads: direct tabulations from the Master File Systems and estimates from administrative record samples. In the first case, survey results are compared to information from what is quite close to a full population tally. This comparison is constrained by the types of data included on the files, which also happen to serve as a sampling frame. In the second case, using administrative records samples, the limita-

tions are the same as those for SOI Studies.

The Internal Revenue Service's Master File Systems are actually accounting operations that are collections of file structures, not simply the Master Files, and the most interesting from our viewpoint are the Transaction Files. Why? Because the better sampling frames are those with many ancillary variables. The Master Files themselves lack that important extra information. Those transaction files offer a means of locating subpopulations that are otherwise difficult to identify. If one seeks information on the upper reaches of the income ladder, then access to these files is critical. Another example is in locating ad hoc operations of short duration, like stock floatation companies.

The value that the tax forms have for surveys also lies in the fact that they can be viewed as a survey instrument for measuring economic conditions. Surveys must often rely on the recollection of the respondent about past events, but the recollection period is quite limited. Some varieties of tax records, though, have an indefinite retention period. Property tax records from the Revolutionary War period still exist in some town halls in New England, for example (although access to these fragile documents is, of course, usually restricted for preservation reasons). More to the point, until very recently, the Internal Revenue Service retained the original Corporation Tax Returns from early in the twentieth century. The raw data on those tax records were sparse indeed compared to more modern versions, but as you can see, this retention renders retrospective studies impossible for interview-based surveys.

Some of these "questionnaires" are fairly simple and direct, but many of the forms contain huge amounts of detailed financial information. And especially, tax information. Since the impact of tax policies is a large issue in economic debates, such data are critical. So how do the regulatory constraints figure in?

We will comment on a few of the tax forms and the associated processing systems with regard to these statistical uses. We will not address excise taxes, nor processing systems for internal operations, nor those about which the author has scant knowledge. Since interest is greatest on individuals and businesses, we will restrict our comments to these areas. We first address the Individual Income Tax Returns and the Individual Master File, although this system is not one we are intimately familiar with. However, this set of records is one that most people are aware of and have experienced. Later, we will focus on Corporations and Partnerships, which are processed through the Business Master File System.

■ Individual Tax Returns

The Individual Master File very nearly covers the entire population. In fact, there is some over-coverage, in part because tax liability and the possibility of amended filings do not end with death. As a sampling frame or source of population data, though, this particular file is not very useful from the statistician's viewpoint. The Statistics of Income Studies, as a result, use transaction files due to the richness of the data there. This comes at a price, however, as those on welfare, dependents, or, in many cases, the retired, for example, simply need not file. Still, this should not be a problem in seeking information on earnings and the effects of tax policies on society.

There are exceptions, though. Certain individuals have their records handled with some discretion, and others who report a tad too much on certain lines of the forms have a simple notation in their computer records indicating that they had complied with the law. The actual transactions, in these cases, are essentially no more informative than the Master File records. We cannot elaborate on exactly which fields they are or who might be affected, though the number is small, for these are confidential matters. External users who are permitted access to the transaction files may have problems with this policy.

The timing issue is the real problem: typically, a survey can only wait so long for the transaction file to be complete. If everyone filed by the April 15 deadline, then that population would likely be complete in mid-year. But a six-month extension to the filing deadline is easy to obtain. Further extensions are granted (though less routinely), and the courts can also step in. Moreover, since the penalty for failure to file is a function of how much you owe, if you do not file and are owed a refund, the regulations do not assess a penalty, in effect

granting unlimited delay. In 1997, an estimated 2.2 percent of all returns filed were from a previous year, and 0.7 percent were more than 2 years past due.

Now, the electronic record that does get created has just about everything reported on the paper forms for a large proportion of the population. Indeed, the basic record allows for over a thousand fields. This allows significant analysis of these records, to the extent that, for many purposes, a sample can be considered as having no nonrespondents. But details are missing for a large number of attachments.

For example, there is no limit on the number of businesses that might be reported, but the transaction record has details on only the largest two, with all others combined onto a third (the SOI Studies capture the first five and create a summary total schedule). For farms, the transaction covers only the first (SOI does two) and so on. In some cases, nearly all of the data are available, but, usually, the interesting details are not on the transaction record (hence, the reason for the extended SOI studies editing procedures).

The individual tax form has many variants, ranging from a simple telephone entry type, through complicated packages of attached schedules. For the most part, however, these are similar in practice to survey questionnaires that use screening options for various portions of the interview. If some threshold is not met, such as having total deductions greater than the standard deduction, then no responses are required for a particular subset of questions. However, unlike the survey instrument, the respondents in our case may decide to provide their own versions of the form, which complicates the data abstraction.

■ Corporation Tax Returns

A corporation is a creature of a State, not the Federal Government. This means that the conditions for its operation can be regulated by the State, including defining the fiscal reporting cycle and various accounting practices.

The different accounting rules covering the various types of corporations have led to the creation of 16 different forms for reporting activities. Unlike a survey form, though, not all of the design is in the hands of an analyst. Five of the corporation tax forms are of little interest for our studies, as they affect operations such as trust funds for decommissioning nuclear plants (Form 1120-ND) or political committees (Form 1120-POL), but others are key in estimating economic indices such as the Gross Domestic Product. The types of forms included in the Internal Revenue Service's Statistics of Income Corporation studies are shown in the table below, along with their Tax Year 1996 estimated populations.

Table 1: Tax Year 1996 Corporations

Form	Volume	Description
1120-S	2,420,000	Subchapter S businesses
1120	2,230,000	The basic Corporation form
1120-A	289,000	The basic form, simplified
1120-F	12,000	Foreign businesses
1120-RIC	8,700	Regulated Investment Companies
1120-PC	3,100	Property & Casualty Insurance
1120-L	1,600	Life Insurance Companies
1120-REIT	500	Real Estate Investment Trusts

The Subchapter S Corporations are businesses that are restricted to having a limited number of shareholders (all must be "natural persons," meaning no corporation or partnership as an owner) where each is taxed directly on his or her share of the business's income. In recent years, the population of the Subchapter S Corporations has been outgrowing the rest of this area. The expansion of this segment of the corporation universe is due to regulatory changes out of both Congress and the Service. In 1996, Congress raised the limit on the number of owners that an S Corporation could have (up to 75 from 35). The change did not immediately boost the number of firms electing this status. An increase in the individual income tax rates, though, may have affected the growth of this type of firm.

Along with the Partnerships, these Subchapter S businesses were used as a vehicle for sheltering income from taxes, though the limit on the number of owners made them less popular. As a result, they are also re-

quired to report active business incomes from activities such as sales separately from "passive" (or portfolio) income. Since the data for these different items are on the paper forms filed with the Service, creation of complete income is possible.

Unfortunately, the S Corporation data on all of the "passive income" items are not included on the transaction records. This means that the sampling frame is without either a "complete" receipts or net income variable (from an economic perspective). On the other hand, since the Statistics of Income Corporation study design places far more weight on the stratification by asset size, the effect on this study is minimal.

We noted that individual tax records had over a thousand fields. Yet we just cited the lack of key data amounts for the corporations. The table below shows just how restricted the items are for stratification.

Table 2: Monetary Fields on the Corporation Transaction Records

Record type	All amount fields	Nontax amounts
"S" Corporations	46	24
Basic Corporation Record	111	51
Real Estate Investment and Regular Investment Corporations	93	37
All Other Corporation Types	47	17

The category "All Other" includes foreign and insurance companies, and, more significantly, consolidated corporations and those with over $10,000,000 in Total Assets. Though these last are among the largest businesses in America, yet only a handful of fields are abstracted. Why?

A major use of the IRS's records systems is to determine which records will be the subject of an audit review. In the case of these large businesses, though, such review is far more frequent than with smaller corporations, as Table 3 shows.

Table 3: Examination Coverage, Fiscal Year 1997

Size of Assets	Returns Filed	Examined	Percent Covered
No Balance Sheet	304,700	3,552	1.2
Under $250,000	1,587,000	18,846	1.2
$250,000 < $1 Mil.	431,500	15,202	3.5
$1 Mil. < $5 Mil.	183,900	14,302	7.8
$5 Mil. < $10 Mil.	27,600	4,421	16.0
$10 Mil. < $50 Mil.	30,500	6,129	20.1
$50 Mil. < $100 Mil.	7,900	1,548	19.6
$100 Mil. < $250 Mil.	7,200	1,647	22.9
$250 Mil. or More	7,800	3,648	46.8

This table, derived from IRS's *1997 Data Book* (IRS 1998), does not include forms from foreign and S Corporations. Yet it is clear that adding more data would not affect the decision on which large businesses are reviewed. Indeed, one might conclude that some firms are under effectively constant review due to their expansiveness.

Then, there is the sheer size of some of the records themselves. The *Wall Street Journal* reported on July 21, 1999, that Citicorp was planning to file its return for last year in the next few months, and that it would have more than 30,000 pages. And this is not the largest administrative record.

On the other hand, because these operations are so important to the economy as a whole, and to our sponsors, the Statistics of Income Corporation Study includes all these businesses if their total assets exceed $10,000,000.

■ Partnership Returns

A partnership is a group of entities (people and/or firms and/or other organizations) that are in some sort of operations together. Yet some joint operating agreements that an ordinary person would consider partnerships are not under the laws of the various States involved or the Federal Government. There are also firms organized as "Publicly Traded Partnerships" which have interests bought and sold on the open market, just like corporations. Conceptually, then, there might be a small

gap in the frame. However, since there is no consolidated filing required, you do tend to get the full legal partnership population.

We will not go into the Active and Passive Income definitions, except to note that, for most of these businesses, the economic receipts and net income can be recreated from the data on the transaction record. The problem area with the capturing of income information lies in the Investment Clubs, where some important fields are not present.

Instead, let us focus on an area that affected both Statistics of Income estimates and the sample design.

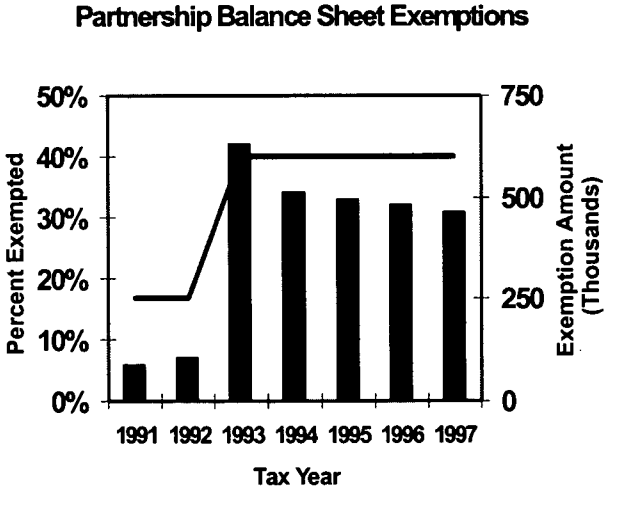
This is about the exemption from an assets reporting requirement. At first, it affected only an insignificant portion of the population, and it was highly constrained. Not only did the Receipts and Assets have to be under $100,000, but there were conditions about who the partners were and the nature of certain kinds of financing. In any case, it presented a minor sample design problem and introduced only a very small bias.

Our initial attempt to deal with this exemption in the design was to set aside a series of sampling classes for records with no assets reported. Since there was no way to tell whether the company was eligible for the exemption on the frame (or on the paper record either that first year), those records were mixed with others of very different character (such as nascent or inactive firms). The resultant heterogeneity was evident in the increased variance for those classes, which in turn forced the sample sizes in those strata beyond our expectations and budget.

So, we did what any good statistical operation would in the place of missing data: we turned to imputation. The models we used were based on regressions through the origin, but were restricted, of course, to information likely to be present on the Business Master File records (McMahon et. al, 1990). We decided against the intercept regression model because, after all, it only made sense that firms without earnings would have few assets, but there was also a strong tendency for the intercept models to start well above the regulatory exemption ceiling (then $250,000). Moreover, we were willing to tolerate a significant amount of noise as long as the return was placed within the likely asset category.

This approach worked well enough for 1989 and 1990, even though the rule had been simplified. But the population claiming the exemption doubled, and the regression coefficients were not stable enough to maintain this design. So, with 7 percent of the population no longer having to report asset information, we were forced, yet again, to redesign the strata (McMahon, 1993).

In this case, for the Tax Year 1991 study, we tried to incorporate the exemption boundaries as strata limits. We are currently re-evaluating this design since the restrictions on this exemption were once again raised. Starting in Tax Year 1993, the requirements are that the firm's business receipts be less than $250,000 and total assets below $600,000 (with the other assorted conditions dropped). This is the definition still in place today. Now, the non-reporting segment jumped to 42 percent (though with the growth in the number of large firms, it has declined to 32 percent more recently). We are currently looking into the bias issues here, and no, we have not imputed values for these records. But as you see, stratification on the size of Total Assets is clearly compromised.



Clearly, these reporting exemptions have an impact beyond that of the sample strata and the SOI tabulations.

Any attempt to validate survey results against administrative records with these constraints will provide only partial verification. For studies that rely on these data, however, we have an implicit bias for estimates of the balance sheet items and their covariates. The question is still unresolved as to whether the relative biases for these variables are constant.

There is one additional issue that must be addressed here, although it covers all sorts of forms, and that has to do with a rule of common law dealing with what a reasonable person would perceive as being compliant. It seems that the government can only be so insistent on the form used for filing tax returns (unlike the sponsor of a mail or interview survey). The author originally became aware of this when the Service first tried to mechanically scan returns. It turned out that very close facsimiles could confuse the devices, yet they were, legally, in full compliance with the regulations. As a result of this "reasonable man" rule, we have situations where "respondent" designed forms and schedules must be accepted. In and of itself, this is not a large problem, but, for pipeline revenue processing, it generates a data input rule that sometimes bypasses information we need for our surveys or will act the same as missing data in any tabulations directly from the Transaction Files.

■ Conclusion

Although there has been a considerable improvement in the quality of the data input to the Internal Revenue Service's Master File Systems, the warnings of decades past must still be heeded. The problem is not the quality alone, nor the incentives for taxpayers to be creative in designing their own forms, but that the various levels of the regulatory structure prevent the data from being as useful as a survey statistician might like.

There are many levels to this structure, some arising out of rules set by the various States, common law, and Federal law, and still more from internal processing rules. At each stage, there are exemptions and exceptions made either to assure some measure of "fairness," relieve burden, or streamline the flow of information. In no small part, it is this factor that causes the datasets to be less useful and complete.

We do not expect any change in this situation. Rather, in many cases, we can expect the intensification of this exemption process, particularly among low-income individuals or favored types of organizations.

■ References

Federal Committee on Statistical Methodology (1980), "Statistical Policy Working Paper 6, Report on Statistical Uses of Administrative Records," U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Internal Revenue Service (1998), "Internal Revenue Service Data Book 1997," *Publication 55B (Rev. 11-1998)*, U.S. Government Printing Office.

McMahon, P. (1991), "Statistics of Income Partnership Studies: Sampling Plan Redesign II," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, P. (1993), "Statistics of Income Partnership Studies: Evaluation of the Revised Sampling Plan," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, P.; Collins, R.; and O'Conor, K. (1990), "Revising the Statistics of Income Partnership Sampling Plan," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.