Statistical Information from Administrative Records in the Federal Tax System

Tom Petska, Internal Revenue Service

The tax return information system is designed to facilitate administration of the Federal tax laws. Under this system, taxpayers report their financial activities, calculate their tax liabilities, and forward a remittance or apply for a refund from the IRS. Although this means of compiling information differs from that of survey-based statistical organizations, the Statistics of Income (SOI) function has a similar mission--to collect and process data so that they become meaningful information and to disseminate this information to its customers and users [1].

Tax returns are filed and administratively processed at IRS regional sites, and selected information from all returns is compiled in a computerized Master File System. Statistics comprising the SOI studies are generally based on stratified probability samples of tax or information returns. As returns are processed into the Master File System, they are sampled based on such criteria as gross income, assets, profits, industry, or presence of supplemental forms and schedules. The sampled returns are then earmarked for additional data abstraction and editing.

While SOI projects are loosely referred to as "administrative records studies," they are clearly more than that--they are, in fact, hybrids of information from administrative and statistical processing. This paper reports on some experiences of using administrative data for analytical and statistical purposes.

In the first section of the paper, a history and operational overview of the SOI function are outlined. The second section highlights differences between IRS Master File compliance data and the statistical information in SOI studies. The third section examines several case studies in which true administrative data, without benefit of supplemental statistical processing, were the primary source of information for research purposes. In the last section, some conclusions about the usefulness of administrative data are presented.

◆ Overview of IRS Tax Statistics Operations

This section briefly describes the history, customers, and operations of the SOI statistical system.

Origins of the SOI Function--The modern U.S. income tax was enacted in 1913 with the passage of the sixteenth amendment to the U.S. Constitution. Subsequently, the Revenue Act of 1916 required the annual publication of statistics, establishing a role for the SOI function. Despite many revisions to the tax law, the original requirement of that Act continues today. Specifically, the current Internal Revenue Code states that "the Secretary (of the Treasury) shall prepare and publish not less than annually statistics reasonably available with respect to the operations of the internal revenue laws."

The SOI Division conducts statistical studies on the operations of the tax laws with respect to individuals, corporations, partnerships, sole proprietorships, estates, nonprofit organizations, and trusts, as well as specialized studies covering both inbound and outbound international activities.

The annual budget of the SOI program is currently about $40 million. While this amount represents a very small portion of the total IRS resources, SOI is a leading Federal statistical organization. SOI's budget covers designing the various projects; overseeing and participating in the statistical processing necessary to accomplish them; and tabulating, documenting, publishing, or otherwise disseminating results.

Customers of SOI Data--The mandate for the SOI program is a responsibility required of the Secretary of the Treasury, and SOI's primary customer is the Treasury Department's Office of Tax Analysis (OTA). Another primary customer is OTA's legislative counterpart, the Congressional Joint Committee on Taxation (JCT).

OTA and JCT use the microdata files produced by SOI as their primary source of information for analysis. In both agencies, microsimulation modeling is employed using SOI data as the primary data base for tax policy analysis and revenue projections. The SOI data are also sometimes matched with other data to build comprehensive data bases that can be used in estimating the overall impact of tax law changes and their effects on tax collections.

Although the bulk of SOI's resources are focused on the statistical needs of OTA and JCT, SOI has many more customers. The Department of Commerce's Bureau of Economic Analysis (BEA) is a significant user of SOI data for estimating components in the National Income and Product Accounts related to individuals, corporations, partnerships, and sole proprietorships. The Census Bureau, also in the Department of Commerce, is another significant data user; however, its needs (unlike those of OTA, JCT, and BEA) are primarily met using IRS Master File "population" data, rather than data from the lower-volume, but content-rich, SOI samples. Other parts of the IRS are customers for SOI data and use it in a wide variety of analyses. Finally, the general public (including academics, "think tanks," accounting firms, and interested citizens) is another important customer of SOI.

SOI Products and Services--Statistics of Income information is made publicly available through both printed publications and electronic media. The *Statistics of Income (SOI) Bulletin* is published quarterly, with each issue containing four to eight articles and data releases of recently completed studies, as well as historical tables covering a variety of subject matter, from Treasury Department tax collections to taxpayer assistance and tax return projections [2]. SOI also produces separate annual "complete reports" on individual and corporation income tax returns, which contain more comprehensive data than those published in the *Bulletin* [3,4]. The *Corporation Source Book* is also published annually, presenting detailed income statement, balance sheet, and tax data by industry and asset size [5].

Periodically, special compendiums of research and analysis, covering such topics as nonprofit organizations, estate taxation and personal wealth, international business activities, and partnerships, are produced. Research articles documenting technological and methodological changes in SOI programs and other related statistical uses of administrative records are also published in a series of reports [6].

The IRS World Wide Web site provides users an easy option for accessing SOI data [7]. At present, 68,000 files are downloaded monthly from the Tax_Stats portion of this site. While SOI manages Tax_Stats in order to make tax-related data on individuals, corporations, and other entities available to the public, we are also the conduit for releasing other IRS information, including the Internal Revenue Service *Data Book* (containing tax collections and other tax administration data), tax return projections, and microdata records of exempt organizations [8].

Statistical Operations--This section highlights statistical operations and procedures utilized in the development and implementation of many of the SOI statistical studies.

- **Sample Design and Selection**--U.S. tax returns are filed and administratively processed at one of ten IRS regional sites, called "submission processing centers." Once processed, IRS compiles selected information from most return forms into a computerized Master File System, which is the informational backbone of the agency. Most SOI operations begin by sampling returns from the Master File System; the Master File offers a sampling frame that enables use of sophisticated and efficient sample designs.

  Statistics compiled for the SOI studies are generally based on stratified probability samples of tax or information returns. As returns are processed into the Master File System, they are assigned to sampling classes (strata), based on criteria such as size of income or assets (or other measures of economic importance), industrial activity, accounting period, or the presence of certain supplemental forms or schedules.

  Each taxpayer, whether an individual or a business, has a unique number, the Social Security Number (SSN) for individuals or the Employer

Identification Number (EIN) for businesses. These unique Taxpayer Identification Numbers (TIN's) are used as the seed for a pseudo-random number (using an algorithm that is a transform of the TIN), which, along with the sampling strata, determines whether a given return is to be selected for the SOI sample [9]. The probability of a return being designated for the SOI sample depends on the sampling rate prescribed for its sample class or stratum and may range from a fraction of 1 percent to 100 percent.

- **Data Capture Techniques**--After sampling, the relatively few data items pulled electronically from the Master File System are substantially augmented with additional items key-entered from hardcopies of taxpayers' returns. Statistical abstracting can take as little as a few minutes for a simple return, to as long as several days for a large corporate return.

  SOI has built a network of midrange enterprise servers in selected IRS submission processing centers that are dedicated to SOI statistical processing. The processing system uses online transaction processing, so that all data capture operations are completed in a single pass. One editor is responsible for ensuring the validity of all data processing for a given return.

  Due to substantial penalties for misreporting, the income and expenditure data reported on tax returns have proven to be more reliable than comparable survey data. Even so, SOI employees go to great lengths to protect against nonsampling errors, such as those due to taxpayer reporting variations or inconsistencies, or data processing errors. In order that final statistics are consistent and reliable, SOI economists develop extensive online tests and error resolution procedures that are applied to each sampled return. The tests and correction procedures are based on the structure of the tax laws and forms, generally accepted accounting principles, and the improbability of various data combinations.

  Editors in submission processing centers, under the direction of SOI economists, "statistically edit" data items in order to make each sampled return internally consistent. Missing data problems arise, albeit infrequently (under 1 percent of the time). Missing items can be obtained through direct contact with taxpayers, or be estimated through imputations based on other return data, prior-year data for the same taxpayer, or same-year data from a "statistically similar" return.

- **Weighting and Estimation**--As noted above, the probability with which a return is selected for inclusion in an SOI sample depends on the sampling rate prescribed for the stratum in which it is classified. Weights are computed by dividing the population count of returns filed for a given stratum by the count of sample returns for that same stratum. "Weights" are used to adjust for the various sampling rates used--the lower the rate, the larger the weight.

  The data on each return in a stratum are multiplied by the weight assigned for the given stratum. To produce the tabulated estimates, as shown in the *SOI Bulletin* and other publications, weighted data are summed to produce statistical totals.

  Of over 200 million tax returns processed each year for administrative purposes, only about half a million are sampled for the various SOI programs. However, since sampling rates generally increase with increases in the size of financial amounts (such as income or assets), the returns in the samples are, on average, disproportionately larger and more complex than those in the master files, which include the population of returns. Thus, in comparison to IRS administrative processing, which captures 100 percent of the tax returns but with limited item content, SOI programs collectively represent a smaller volume, but with a proportionately higher fraction of complex returns and with greater item content.

◆ Administrative Data for Statistical Purposes

Most of what is familiar to users of SOI data are the data products from traditional SOI sample studies. However, there is another side to SOI, that of a user and developer of data from the IRS Master File System.

As noted above, SOI has, for many years, selected samples from the Master File System, but use of these data as program content is less frequent. In this section, examples of the use of administrative tax data from the U.S. and other systems are described, and commentary is offered on its current value and potential for additional analytical purposes. In IRS terms, such data files are referred to as the Individual or Business Returns Transactions Files (i.e., IRTF and BRTF), key inputs to the Master File System.

SOI's traditional data editing processes are labor-intensive. Few SOI studies have been accomplished without substantial manual data abstraction and editing. However, there is a reluctance to change this, since the final weighted estimates and overall quality of data are so heavily dependent on the accuracy of the data editing process.

In general, the need for manual abstraction and editing is dependent upon two issues:

- ❒ Is the statistical item content (e.g., the financial data) included in the population files adequate for analytical and estimation purposes?
- ❒ Does the level of data complexity adversely impact an acceptable level of data quality for the return type or population being studied?

Concerning item content, the issue is fairly straightforward. If program-critical data were available on the return form or attached schedules or worksheets, but not abstracted into the Master File System, the only alternative would be to obtain copies of these forms and schedules and abstract these data.

On the issue of complexity, the situation is less clear. In the U.S., many relatively low-income individual income returns are quite simple, with only a limited number of income types and other taxpayer-reported characteristics. For such cases, an automated or high-level system of reviewing outliers and imposing error corrections to the sampled file cases would, in all likelihood, yield reasonable results.

But as complexity increases, this may not be an acceptable means of data editing. For example, large U.S. corporation returns, often with multinational operations, are extremely complex, and SOI staff spend hundreds of hours reviewing and correcting these data, even after the initial abstraction and editing have taken place. These often come about as "referrals" to professional staff for closer scrutiny. In extreme cases, where the return complexity is so high or the data reported are incomplete, SOI staff correspond with the taxpayer, which is time-consuming and often a source of delay in project completion. But our view has been that, in such cases, it is essential.

Many statistical agencies in the U.S. and elsewhere have begun to develop automated data edit systems, building an "artificial intelligence" knowledge base for data editing. Most applications have started with fairly simple item content where the editing relationships could be specified to handle the majority of cases. However, before such routines can become the norm for studies with more complex data content, such as most business tax returns, extensive additional study of data relationships and taxpayer reporting tendencies is needed.

◆ Administrative Data in Statistical Studies

This section provides brief descriptions of experiences of using true "administrative data" for statistical or analytical purposes.

Administrative Data in SOI Programs--For many years, SOI has, in addition to sampling, brought in program content data from the Individual and Business Returns Transactions Files. Currently, such data are essential parts of record content for SOI sampled returns in the Individual, Corporation, and Partnership programs.

This process was initiated many years ago from the perspective that it would save resources from re-keying

these data, but it has not been without its problems. Data editing for the Master File System not only differs from SOI statistical processing, but SOI has had very little control in these processes.

Still, with many years of experience in using these data as part of SOI program content, subject-matter analysts have devised ways of making these data useful and of high quality so that the processing efficiencies envisioned years ago have generally been realized.

Subnational Economic Estimates--The SOI samples are not reliable below the national level, which is a conscious decision based on priorities and budgets, since sample sizes would have to be substantially inflated. Thus, subnational estimates require data from the Master File System since they account for the entire population of returns. Presently, State-level estimates are produced for Individuals, Proprietorships, and Partnerships, and ZIP Code data are also produced for Individuals. In general, these programs have been largely successful.

The data on individual income and taxes by State, which are derived from the IRTF data, are published annually in the Spring *SOI Bulletin* and also made available on the SOI Internet site.

Migration Data--In a joint venture with the Census Bureau, IRS Individual Master File data have been appended with geographic codes for use in Census State and county statistics programs. These data have been tabulated by State and county, and net migration has been calculated. These data are made available to the public from SOI's Statistical Information Services (SIS) office.

SOI Fiduciary Study--Recently, a study of Fiduciaries was completed solely from the population data on the filing of Forms 1041, Fiduciary Returns, in the IRS Business Returns Transactions File. This study included the population of returns and developed edit rules to ensure reasonableness and consistency of data at the "micro" level. This study has since been published in the *SOI Bulletin* [10].

Two other case studies in which IRS Master File System data were the primary data source include the following:

IRS Compliance Data Warehouse--The IRS has attempted to develop a centralized compliance data warehouse (CDW) from the Master File System for use in tax compliance studies. The warehouse is a large relational data base consisting of population files from the IRTF and BRTF. The system has been a mixed success, but current plans are under way to improve it by adding more data sets and increasing accessibility.

Census Bureau Demographic and Economic Programs--By law, the Census Bureau has been provided with annual extracts of individual population data from the Master File System. Although item content and uses are restricted, these data have been used effectively in periodic censuses and other statistical programs.

Finally, two other case studies, both in the early developmental stages, are mentioned because of their far-reaching goals of attempting to build a comprehensive statistical data base exclusively from Master File System data.

Statistical Data Warehouse--From the perspective that "to get what you really want, you may have to do it yourself," SOI has taken the initial steps to acquire both the data sets and operational systems to build a data warehouse exclusively for analytical and statistical purposes. Like the CDW, the primary data would consist of inputs from the IRS Master File System. However, since it is being designed exclusively for statistical purposes, improved accessibility and compatibility with statistical software are essential ingredients.

South African Tax Statistics--Current and former SOI staff have had an advisory role with officials in the South African Department of Finance and the South African Revenue Service to develop the capability to do microsimulation modeling of their individual and business tax systems. Since resources are extremely limited, work is under way to ascertain if the population files of all individual and business tax returns could be developed as the exclusive source for statistical analysis and forecasting [11].

◆ Summary and Conclusions

In conclusion, one might ask what do these administrative data studies have in common and what are the factors attributing to their success? Here are some such factors common to many of these studies:

- ❒ Data were needed without adequate resources for transcription and editing.
- ❒ Data were needed with a very short time horizon.
- ❒ Data were needed for the entire population (or something close) because subnational estimates were a primary focus.
- ❒ Longitudinal analysis of individual case behavior was a primary focus, and sample data often did not have a high degree of year-to-year overlap.
- ❒ The administrative data were perceived as a rich and largely untapped resource that could yield substantial benefits if properly developed and used.

Despite SOI's long and successful history of sampling and editing data from the tax compliance "pipeline," there have thus been many instances, both within the IRS and elsewhere, where administrative data have been successfully used for statistical purposes. This paper has cited a few such examples with the hope that lessons can be learned from these successes so that more and better uses can be developed in the future.

◆ Acknowledgments

Portions of this paper have benefited from prior overviews of the SOI function written by Fritz Scheuren and Tom Petska for *Business Economics* (1992); the *Proceedings of the National Tax Association* (1992); and the *Journal of Official Statistics* (1993). Special thanks to James Dalton, Beth Kilss, Mark Mazur, and Dan Trevors for their many helpful comments on an earlier draft of this manuscript, and again to Dan for prepublication processing and electronic submission to ASA. Any errors that remain are the responsibility of the author.

◆ Notes and References

[1] *Principles and Practices for a Federal Statistical Agency* (1992), Martin, Margaret E. and Straff, Miron L. (editors), Committee on National Statistics, National Academy Press.

[2] *Statistics of Income (SOI) Bulletin*, Publication 1136, Internal Revenue Service.

[3] *Statistics of Income--1998, Individual Income Tax Returns,* Publication 1304, Internal Revenue Service.

[4] *Statistics of Income--1998, Corporation Income Tax Returns*, Publication 16, Internal Revenue Service.

[5] *Source Book of Statistics of Income--1998, Corporation Income Tax Returns*, Publication 1053, Internal Revenue Service.

[6] Statistics of Income: *Turning Administrative Systems Into Information Systems--1999,* Publication 1299, Internal Revenue Service.

[7] The SOI Internet website, Tax_Stats, can be accessed at http://www.irs.gov/tax_stats.

[8] *2000 Data Book*, Publication 55B, Internal Revenue Service.

[9] The algorithm used for generating the Taxpayer Identification Number (TIN) transform generally stays the same from year to year. This longitudinal character of the sample design improves the estimates of change from one year to the next.

[10] Mikow, Jacob M., "Fiduciary Income Tax Returns," *Statistics of Income (SOI) Bulletin*, Winter 2000-2001, Volume 20, Number 3, Publication 1136, Internal Revenue Service.

[11] Petska, Tom, "Exporting a Statistical System: Towards Establishing a Tax Statistics Function in South Africa," *2000 Proceedings of the American Statistical Association, Sections on Social and Government Statistics.* ■