Data Interpretation across Sources: A Study of Form 990-PF Information Collected from Multiple Databases

Melissa Ludlum, Statistics of Income, Internal Revenue Service

Statistics of Income Division RAS:S:SS:S, P.O. Box 2608, Washington DC 20013-2608

Key Words: Private foundation, stratified sample, administrative data, coefficient of variation

I. Introduction

Private foundations contribute billions of dollars each year to charitable initiatives directed toward such issues as environmental protection, health and human services, promotion of the arts and humanities, and educational outreach and opportunities. With several hundred billion dollars in asset holdings, private foundations constitute a substantial segment of the nonprofit sector. Unlike public charities, which are often funded by, and therefore directly accountable to, the public, private foundations generally receive funding from a limited number of sources. Furthermore, an individual or small group typically controls the majority of a foundation's activities. Due to this narrow base of support and control, detailed financial information on private foundations is often more difficult to obtain than similar information for other charitable organizations. In many cases, data collected from tax return records and disseminated by the Internal Revenue Service (IRS) provide the most comprehensive information available on the financial composition and charitable giving habits of private foundations. Statistics derived from these sources can provide a window into the charitable activities of these organizations. Additionally, the information supplied to IRS provides insight into both the investment portfolios of private foundations and into the nature and amount of their charitable and noncharitable expenditures. These data can also reveal emerging trends and developments in the private foundation segment of the nonprofit sector. Analyses conducted using such data provide a framework for the development of tax policy related to private foundations and assist practitioners and foundation staffs in the establishment of key self-governance principles.

Unlike the majority of taxpayers, who report information to IRS on "tax returns" designed to assist in the calculation and payment of income taxes, private foundations complete "information returns" designed to collect a wide range of information. Because of their primarily charitable missions, private foundations receive exemption from Federal income taxes; they are, however, subject to an array of stringent legal requirements. Under regulation, they are required to distribute a certain percentage of their asset holdings to charitable activities each year. Secondly, although private foundations are exempt from *income* tax, they are required to pay an *excise* tax on their investment income. In addition, unlike corporate or individual taxpayers, private foundations are subject to public inspection requirements. This means they are responsible for ensuring that their annual information returns, known as Forms 990-PF, are widely available to the public. Each year, private foundations file the extensive, twelve-page return with IRS, reporting standard income statement and balance sheet items, as well as additional information on charitable distributions, compliance with rules that govern private foundations, involvement in various types of activities, and certain employment information.

The public inspection requirement promotes increased data availability and thus provides a wide range of analysis opportunities for interested researchers. Users can obtain micro-level data from Forms 990-PF from a number of sources. For example, independent organizations such as the Foundation Center and GuideStar obtain Forms 990-PF from IRS and post them to the Internet on a continuing basis. Another organization, the National Center for Charitable Statistics (NCCS), makes an annual file of return data from the IRS Returns Transaction File (RTF) available to researchers wishing to obtain data for large numbers of

organizations. This file, which the IRS provides to the NCCS annually, includes limited data for the population of Form 990-PF filers. The Statistics of Income (SOI) file provides yet another resource for private foundation data. This file includes error-corrected data items for a sample of Forms 990-PF.

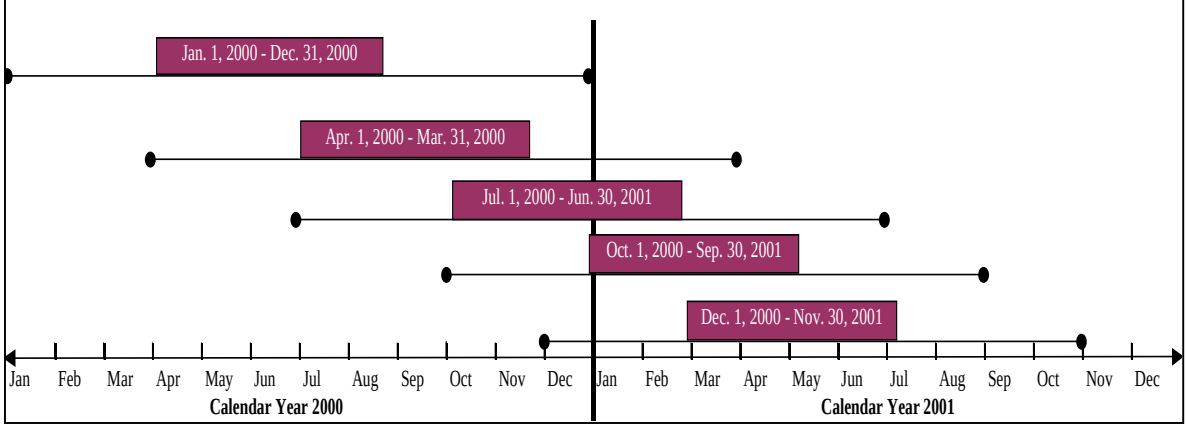
While the numerous available data sources enhance research options, reconciling them to one another can be a difficult experience for data users. Measuring data quality and discrepancies among them is a formidable, but necessary, challenge. Before conducting analysis, researchers should be aware of the range of available data sources, as well as the limitations and advantages that characterize the data sets obtained from these sources. Such information is especially important when supplementing data from any one source with information obtained from another. Understanding the unique characteristics of data obtained from each source also helps to explain, and reduce, statistical variation between them. Additionally, assessment of these data sources allows opportunities to combine information from them, possibly reducing data collection costs and expediting processes. This paper will discuss two IRS-derived data sources, the IRS Returns Transaction File and the SOI File, and determine the various quality and consistency issues associated with each source. It will describe the various administrative data sources from which private foundation data may be obtained, outline the methodology for identifying comparable tax returns to create a standardized dataset, examine the results of preliminary analysis conducted on aggregate and micro-level statistics from the datasets, and present conclusions and future applications derived from the research conducted.

I. Data Sources Overview

When IRS receives a Form 990-PF, a limited number of data items are key-entered as the return is processed and posted to what is known as the RTF. IRS creates an annual RTF extract, which includes information from all returns received by IRS during a given "processing," or calendar, year. The extract includes approximately 100 money amounts, or financial items, with an additional 85 fields of codes and other non-financial information. When working with RTF-derived data, it is important that users are aware that the file may include a number of superfluous records, such as duplicate or incorrectly filed returns. Under most circumstances, data users should remove such records before conducting most analyses.

When using RTF data, several important factors should be taken into account, particularly if the data are used in conjunction with data from other sources. First, the timeframe that a set of returns represents must be considered. An extract for a given calendar year should include the "population" of Forms 990-PF filed with IRS during that year. However, organizations file Form 990-PF based on reporting year, which corresponds to the year actually printed on the return. As illustrated by Figure A, which shows examples of accounting period that can be present in a typical Reporting Year, an organization determines its reporting year based on its accounting period, specifically, based on the month in which its accounting period begins. Thus, an organization would file a Reporting Year 2000 return if its fiscal year

Figure A: Examples of Filing Periods Represented by Returns Filed for Reporting Year 2000



accounting period *began* in any month of Calendar Year 2000 [1]. However, many Reporting Year 2000 returns, such as those with accounting periods that began in December 2000 and ended in November 2001, would not have posted to the RTF until Calendar Year 2002. When conducting time-series analysis, or analysis among multiple data sources, it is important to understand the relationship between accounting periods, calendar or processing years, and reporting years in order to achieve the most consistent dataset possible.

Secondly, although different types of organizations file the same return, they may not necessarily be subject to the same tax treatment. Both tax-exempt private foundations and nonexempt charitable trusts are subject to the private foundation rules and are thus required to file Form 990-PF. However, in some cases, nonexempt charitable trusts may also be responsible for paying income tax, reported on a separate, additional return. Such a distinction could easily affect the behaviors of these organizations. Therefore, these segments of filers should be identified and treated as distinct types of entities, thus allowing the opportunity to examine these data in both separate and aggregate frameworks. If an RTF data user is aware of this distinction, he or she can easily identify nonexempt charitable trusts and private foundations based on their assigned subsection codes.

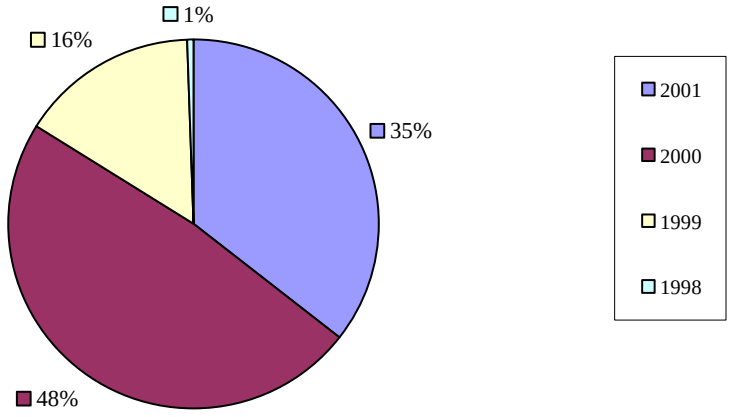
Based on postings to the RTF, SOI samples approximately 10 percent of all Forms 990-PF filed for a given reporting year. The SOI file contains more than 200 financial items, with 75 fields dedicated to codes or non-financial information. The SOI staff enters data into an online system, which identifies taxpayer and other errors, which are corrected during the data entry process. Often, supplemental information is included with Forms 990-PF on schedules and other attachments. Where appropriate, information from these attachments is used to supplement or enhance data reported by the filer. A typical completed reporting year sample includes numerous allocations. For example, SOI made nearly 17,000 allocations for the Reporting Year 2000 sample.

Unlike the RTF extract, which includes all returns filed in a given calendar year, the SOI Reporting Year sample must be conducted over 2 calendar years. This method of data collection is used as it ensures almost complete coverage of a reporting year population, preventing organizations from being excluded from the sample in cases where their returns are filed outside of the anticipated calendar year. Like the RTF, the SOI file includes returns filed by nonexempt charitable trusts, but duplicate returns and returns with inconsistencies that cannot be resolved are removed before dissemination.

II. Analysis Methodology

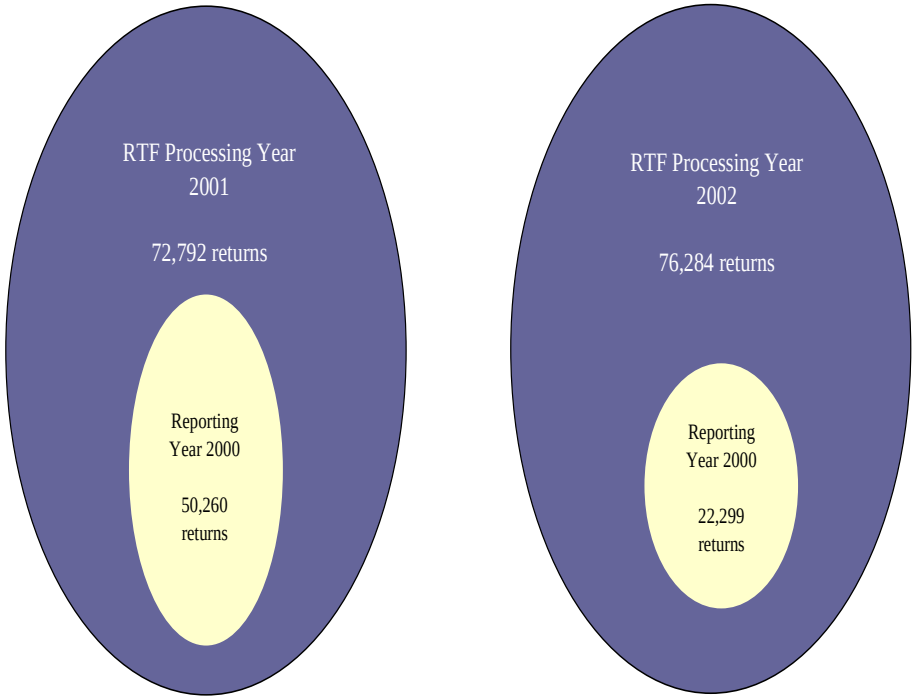
The first challenge in measuring consistency and quality issues between the two sources was to standardize and combine the data sources by creating a standardized dataset; the resulting

Figure B: Reporting Years Represented in the Combined Extract



dataset was designed to include data from a single reporting year and to be free of duplicate and extraneous records. To create the standardized dataset, a series of steps was taken to ensure that the highest possible level of consistency was achieved between RTF and SOI data.

Figure C: Components of the Combined Returns Transaction File



The analysis includes returns filed for Reporting Year 2000, which IRS received over several calendar years [2]. To identify the appropriate returns, while still limiting the number of years of RTF data that were included in the analysis, the final dataset was limited to those extracts containing returns posted in Calendar Years 2001 and 2002. This timeframe coincides with the period in which data were collected for the SOI Reporting Year 2000 file.

In addition to including nearly the entire population of timely-filed Reporting Year 2000 Forms 990-PF, the combined extract also included returns filed for other reporting years between 1998 and 2001. Figure B shows the percentage of returns from each reporting year that appeared on the 2001 and 2002 combined RTF extract. Less than half of the returns on the extract represented Reporting Year 2000, and a substantial number were filed for Reporting Year 2001, with smaller but significant numbers representing other reporting years.

Figure C illustrates the number of returns associated with each year in the combined extract. The calendar year populations appear in the larger ovals, with the Reporting Year 2000 subset represented by the smaller ovals. Only the 72,559 returns filed for Reporting Year 2000, identified as the sum of the subsets of the 2 calendar years, and represented in the smaller ovals, were initially considered for inclusion in these analyses.

Once the subset of included organizations was narrowed, based on reporting year, several additional steps were taken to arrive at a standardized dataset. Records were removed if their associated “status codes” indicated that the organizations were inactive or no longer exempt. In some cases, returns appeared more than once on the RTF. A series of procedures removed these duplicate returns from the standardized dataset. Finally, the completed dataset included only returns filed by private foundations, identified based on the assigned subsection code. Once concluded, these steps revealed an RTF population of 68,355 returns suitable for inclusion in the analysis.

For comparison purposes, the SOI file for Reporting Year 2000 was used for this analysis. The sampling period for the file began in January 2001 and continued through December 2002. The file is a random Bernoulli sample, based on organization type and asset size, using different parameters for private foundations than for charitable trusts. In addition to being subject to different tax treatment than private foundations, nonexempt charitable trusts are generally much smaller, in terms of asset size, than are their tax-exempt counterparts. Private foundations with $10 million or more in assets and nonexempt charitable trusts with $1 million or more in assets were selected at rates of 100 percent, with decreasing rates applied to smaller-sized organizations [3]. For the initial research, the SOI file remained largely intact, with one exception: all returns that were ultimately determined to be "charitable trusts" were removed from the data. While returns filed for charitable trusts were removed from the RTF based on subsection code, they were removed from the SOI file using a more perfected data field, which is not available on the RTF [4]. This field rectifies errors in organization type that are often present on the RTF at the time of sampling.

IV. Aggregate Analysis

After standardization of the data sets, aggregate RTF and SOI data were compared. For major data items, the two sources did not provide significantly different results. Figure D is a comparison between the coefficients of variation, used to estimate of SOI sampling error, that were calculated for three major data items, and the percentage differences between estimates derived from the RTF and SOI data files. Note that, for two of the three categories, total revenue and total expenses, the percentage difference between the two datasets falls inside of the sampling error estimates. For one category, fair market value of total assets, the difference by which the RTF amount exceeds the SOI amount is somewhat larger than the sampling error. The larger difference may be attributed to a variety of differences in editing and error correction, which are driven by the purposes for which the data are collected. While RTF data entry operators often key data directly from the Form 990-PF for examination and tax collection purposes, SOI editors may substitute amounts from attachments in lieu of amounts reported on the return. These types of substitutions and corrections allow SOI to produce statistics that are more accurate and to provide additional data items for customers that use microdata files.

Figure D: RTF and SOI File Comparison: Percentage Differences and Coefficients of Variation

Item	Coefficients of variation (percentages)	Difference RTF to SOI (percentages)
Total assets (fair market value)	0.66	4.83
Total revenue	1.50	0.65
Total expenses	2.84	2.19

V. Microdata Analysis

To analyze microdata fields between the two datasets, individual returns were linked from the SOI file back to the parent RTF, based on their unique taxpayer identification numbers. Returns were not linked unless they appeared on the RTF dataset that was used for aggregate analysis. Once linked, the files were compared for inconsistencies between major data items. The inconsistent fields were then weighted, using the SOI design-based weights, to determine the effects of the SOI correction processes on the overall population estimates. A field was identified as "inconsistent" if the amount transcribed to the SOI file differed by more than $25 from the amount that appeared on the RTF. While corrections were made to many data items common to the two datasets, nine major fields appeared to be corrected by SOI editors most frequently.

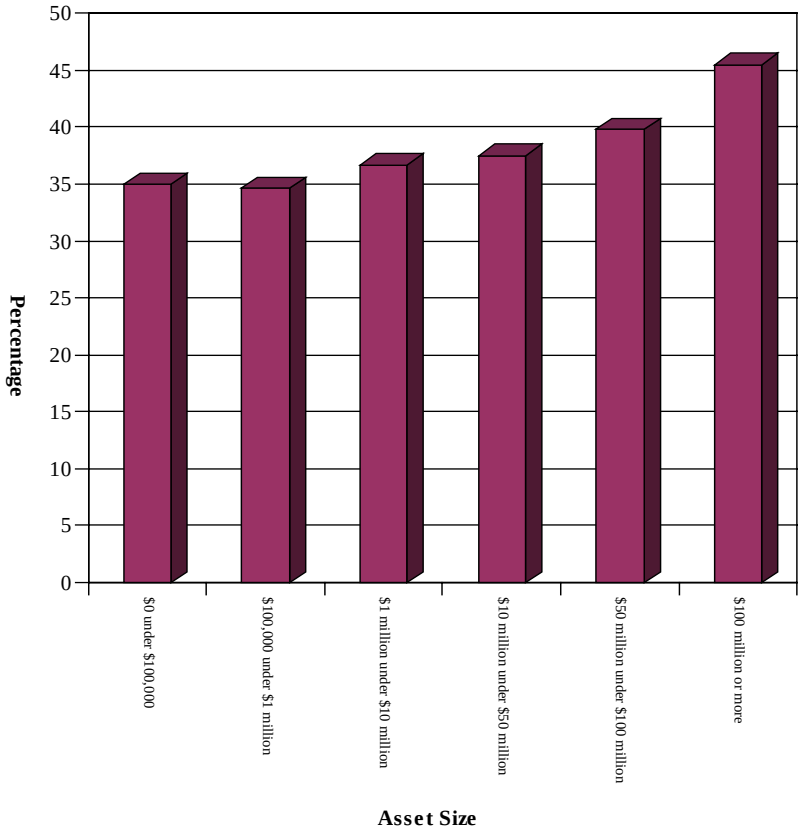
The three balance sheet items that represented securities--corporate stock, corporate bonds, and Government obligations--were corrected most often and, based on the median values of these corrections, with the most magnitude. Figure E shows RTF fields to which SOI editors commonly made corrections. In most cases, these corrections probably resulted from procedural differences in data entry, rather than operator error. SOI data entry operators collect information from supplemental attachments and schedules, in addition to the data that appear on the Form 990-PF, to enhance the quality and accuracy of the microdata. The maximum and minimum correction values exemplify the effects of large keying errors on the RTF. Weights associated with the returns identified as corrected were applied to estimate the effects of SOI data entry on the overall population of private foundations. The

Figure E: Unweighted and Weighted Corrections, Amounts and Values

Data Item	Minimum value	Maximum value	Unweighted Corrections			Weighted Corrections		
			Number	Median value	Mean value	Number	Median value	Mean value
Corporate stock	-5,241,441,621	118,170,705	1,640	-986,193	-16,639,345	10,725	-70,299	-2,763,242
Corporate bonds	-186,930,409	441,778,508	657	-984,966	-3,264,243	3,974	-123,775	-715,434
Government obligations	-344,684,265	454,418,685	538	-425,973	-1,615,332	3,354	-58,743	-270,657
Total assets, book value	-19,021,602,054	2,276,122,860	367	-21,264	-54,123,269	2,311	-5,159	-8,614,235
Total assets, fair market value	-34,824,317	397,295,763	297	-30,001	1,015,237	2,701	-19,247	-1,101,468
Total expenses	-70,188,315	28,400,237	240	-13,103	-506,701	1,567	-2,679	-81,322
Total revenue	-70,188,315	117,315	237	-15,127	-941,870	1,241	-986	-185,424
Undistributed income	-28,751,786	11,363,248	222	3,664	134,594	4,009	-438	11,105
Other revenue	-70,188,315	291,274	206	-15,033	-593,730	1,106	-1,009	-114,774

categories of stocks, bonds, and Government obligations remained the most-often corrected financial items, after the weights were applied. The category "undistributed income," a field that represents required charitable distributions that foundations did not make in Reporting Year 2000, represents a larger portion of the total weighted corrections made than in the unweighted total. This indicates that more changes to the field were made to smaller, and therefore more heavily weighted, asset-size class returns. The relationship between foundation size and number of corrections was examined by arranging returns included in the microdata analysis into commonly used used "asset-size" categories. Figure F shows the percentage of returns with at least one correction to one of the nine data items examined, by asset size category. The proportion of corrections, generally, increased slightly with foundation size. More than 45 percent of the returns filed by the largest organizations, those with assets of $100 million or more, had a least one correction, indicating that the largest organizations are proportionally

Figure F: Percentage of Returns with at Least One Correction, by Asset-Size Category



more often corrected than are their smaller counterparts. Overall, for the nine selected items, nearly 40 percent of the returns in the SOI sample had data inconsistent with that appearing on the RTF.

VI. Conclusions and Future Research

Based on this research, several important conclusions regarding data consistency, compatibility, and collection can be reached. In the past, SOI has been hesitant to supplement information unavailable on the SOI file with similar data from the RTF. However, it appears that these data can be used as complements, as long as the RTF data files are properly restricted to be consistent with the SOI file. While the SOI dataset is the only source for many data fields, in the future, the RTF may provide a valuable source for obscure, but sometimes necessary, data items. An important conclusion regarding data collection can also be reached based on this research. Currently, only a handful of items, none of which is financial, are incorporated directly from the RTF to the SOI transcription process. In many cases, however, some items that are available on the RTF 990-PF file remain largely unchanged during the SOI editing process. In the future, SOI may wish to build on this information and identify items that can be captured directly from the RTF to reduce the redundancy of operator transcription. SOI resources could then be directed toward transcribing additional data items, which may not currently be available from any source.

Several future research options are available that could also help to illuminate data quality and collection issues. Currently, a sample of large-case returns that are included on both the RTF and SOI file is being transcribed based on information that appears directly on the Internet-posted, publicly available return. The data are being collected without additional information from attachments or schedules being transcribed. The information will provide insight into an avenue that researchers commonly use for information—the Internet, and will determine if the data posted by these organizations are consistent with those collected by IRS. Another valuable venture would involve comparing data from the SOI and RTF files for a number of years to ensure that that RTF data quality does not fluctuate between calendar years. This information could assist in determining definitive sources for specific data items. Ultimately, the results of this research may assist in improving resource allocation in the collection and dissemination of private foundation data.

Notes and References:

[1] For example, a return that had an accounting period that began in January 2000 and ended in December 2000 was filed for Reporting Year 2000. This return would have likely been posted to the RTF in Calendar Year 2001, as the required filing data is five and one-half months after the end of the accounting period.

[2] In some cases, a return that was file late or by a taxpayer that received numerous extensions to file could have been received by IRS outside of the traditional, two-calendar year window.

[3] The realized sampling rates for the Reporting Year 2000 SOI study of private foundations are shown below:

Fair Market Value of Total Assets	Realized Sampling Rate (percentage)
Private Foundations	
Under $125,000	0.3
$125,000 under $400,000	0.8
$400,000 under $1,000,000	1.9
$1,000,000 under $2,500,000	4.3
$2,500,000 under $10,000,000	21.0
$10,000,000 under $25,000,000	100.0
$25,000,000 or more	100.0
Charitable Trusts	
Under $100,000	1.2
$100,000 under $1,000,000	13.4
$1,000,000 or more	100.0

[4] Private foundations and charitable trusts were identified on the RTF based on their respective subsection codes. Private foundations are assigned a subsection code of “03,” while nonexempt charitable trusts are assigned a subsection code of “92.” Generally, organizations were also coded for the SOI File based on their subsection codes. However, in cases where subsection codes appeared to be incorrect or were not available, SOI staff conducted additional research to determine the proper subsection code for organizations on the SOI file.