Disclosure-Proofing the 1996 Individual Tax Return Public-Use File

Peter Sailer, Michael Weber, and William Wong, Internal Revenue Service

Beginning with Tax Year 1960, the Statistics of Income (SOI) Division of the Internal Revenue Service and its predecessor organizations have periodically released public-use files consisting of stratified samples of unidentified individual income tax return data. The purpose is to provide researchers outside of the Government an opportunity to analyze the tax system and model proposed tax law changes.

For 1960, as is true to this day, the Public-Use File was a subsample of the regular Individual Statistics of Income file used to produce the tabulations in our Statistics of Income reports--a subsample with limited item content. (Only 7 codes and 17 money amounts were included for 1960--we did not feel that our customers' computers could handle a larger file.) The only steps taken to deal with disclosure problems for the 1960 file were the elimination of identifying names; identifying numbers (such as SSN's); and all geographic information. We actually included a 100-percent sampling stratum, on the theory that it was needed to make valid estimates at the upper end of the income distribution.

As time progressed, larger computers made it possible to include more data items in the Public-Use File. By the mid-1970's, the file had settled into a format of about 30 codes and 150 money amounts. In response to urging from our users, State codes were added for records with incomes under $200,000 (a decision was made that any geographic information for returns with incomes of $200,000 or more would present a disclosure risk).

◆ Disclosure-Proofing Enhancements for 1985

Research conducted by SOI in the early 1980's indicated that some of the items included in our Public-Use File were also available from other sources in identified form. For example, anybody going to the county courthouse and looking up the valuation of a certain residential property, as well as who owned it, could come close to figuring out what that individual's real estate tax deduction should be; one only needed to multiply the valuation by the county's property tax rate. We found that the salaries of major corporate officers could be looked up in their corporations' annual reports--although, luckily, all of these corporate officers (or their spouses) appeared to have extensive additional salaries from other sources.

As a result of our research, we instituted the following methods of dealing with disclosure problems:

- First, we subsampled the 100-percent stratum to make sure that no user could be sure that a given individual was in our file. Our lowest weight became 3.00.

- Next, we introduced some noise into the file by blurring certain amounts we deemed to be sensitive. For example, after dividing the file into segments by demographic characteristics, we sorted each segment by size of salaries and wages, then computed averages for three returns at a time, and assigned that average to each of the three returns. The advantage of using blurring is that the totals for the various demographic groups still come out exactly right. Blurring was performed on all items which our research had indicated could be obtained from some source in identified form.

- And finally, we rounded all our statistics to four significant digits, so that even the unblurred fields could not be matched exactly to the dollar, should someone have found an unexpected source of matchable data.

These methods for dealing with disclosure are spelled out in detail in a paper by Strudler, Oh, and Scheuren, given at these meetings in 1986 [1].

◆ New Disclosure Research

Beginning in the mid-1990's, we took the additional precaution of examining the outliers for each field and assigning to them a zero-percent probability of being selected in the subsampling process. Three or four senior technicians, including the co-authors of this paper, examine the records that have the largest amounts in each field--as many as 10 records per field, although there is much overlap between the records selected for each of the fields. If the record can be directly identified with one variable, it is obviously out. But any record that has a combination of items that make it unique is also tossed. For example, a record may have a source of income that is twice as large as the second-highest amount for the same source in the same demographic group.

We also decided to search for new sources of data that could be used to break into the file. There have been many papers written recently, suggesting that traditional methods of disclosure-proofing Public-Use Files are no longer adequate in this age of the information explosion [2]. One source of concern is the data bases available to marketers, such as those produced by Dataline America, Donnelly, and InfoUSA. We decided to test the Public-Use File by trying to link it to one of these.

The file we chose for comparison was advertised as having data, among other topics, on total household income, marital status, home mortgage, presence of children, charitable contributions, and age. In other words, it presented six potential items on which to match to the Public-Use File. Luckily, the identified file did not contain actual amounts. Family income was coded in $5,000 increments, and top-coded at $145,000. Marital status was basically a yes/no code (married/not married), as was the presence of children code. Age was available in 1-year increments, but age data on the SOI side were a yes/no code (age 65 and over, or not). Home mortgage was a one-digit code with ten possible values relating to the size of the total loan--SOI data referred to the amount of interest paid in a given year. In the absence of information on interest rates and length of payments, we decided to treat home mortgage as a yes/no code. The charitable contributions code related to the type of charity, whereas SOI has amounts, so it too could only be used for a yes/no comparison. So, we basically had five yes/no comparisons and one comparison (income) on a series of code values.

In order to test how useful the marketing file might be in breaking into the SOI file, we purchased all medium to high-income records for the least-populated State--Wyoming. Because of the small size of the file, we dispensed with any sophisticated matching software. We simply ordered the file by name and address, and then looked for matches in our pre-disclosure-proofed file. There were 123 returns that matched up to a record in the marketing file. We assigned each a score based on whether the marketing data agreed with the SOI data--one income class and five yes/no questions, so a perfect score was a "6." Had all the data on all 123 records agreed, the total score would have been 6 x 123, or 738. The actual total score was 219.

The reason for this dismal (and therefore, for us, heartening) match rate was the overwhelming presence of "U" answers--"U" as in "UNKNOWN"--in the marketing file. The creators of this file had to depend, in large part, on consumers filling out and sending in those little warranty cards received with major purchases, plus responses to various membership and customer "surveys." Obviously, Americans are none too diligent in responding to these inquiries. The only question where (just) over 50 percent of the answers corresponded to the information given to IRS was the age 65 or over question. Marital status corresponded to the IRS data for only 1.6 percent of the respondents.

If this marketing file was all we had to worry about, our disclosure-proofing efforts would represent overkill. But the fact of the matter is, we did not know the quality of every publicly available source. Also, we could not be sure whether we had overlooked some sources of data that could be used for breaking-in purposes. While we had done research back in the early 1980's, it had not been confirmed outside of IRS.

The way we decided to tackle the problem this time was to hire a private detective who makes his living by finding out details of other people's lives. We told him which items were on the Public-Use File, and asked him

whether he could go out and find them for a hypothetical target. We stipulated that we did not want him to break the law, even hypothetically.

This investigation proved beneficial, since our detective found a source for sole proprietorship gross receipts of which we were previously unaware. He also predicted we would find that it was not a very accurate source, since it relied on business persons responding to the request for this information. We did a little research, and confirmed that the data base did exist, could be accessed at a reasonable cost, and, as our detective had predicted, usually contained no information on gross receipts, or what must have been very old information. However, we did find a few records where accurate information was present. So, by our rules, gross receipts had to be added to our list of fields to be blurred.

◆ Testing the File

So, now, we ran our traditional blurring, top-coding, and rounding programs, and eliminated the extreme outliers. The blurring--i.e., averaging--was done within groups of returns with identical demographic characteristics. We felt reasonably confident we had distorted, one by one, those fields that could be used to break into the file. But what about the combination of the distorted fields? Here is where we felt we might miss something, so we turned to Westat, our statistical contractor, to use one of the much-vaunted statistical matching programs to evaluate our file. The program they used was Automatch [3].

Here, we faced a difficult choice: if Westat was going to attempt a statistical break-in, we had to provide them with some data with which to do it. What we had available was IRS's Master File of all individual income tax returns. We knew that no outside intruder would have data that were as close to what was on our file; however, we also knew we would never be able to purchase each available data base that had matching data--let alone be sure we had purchased the best one possible.

So, we gave Westat records for all of the highest-income taxpayers on the Master File--those represented in the 1-in-3 sample in our Public-Use File--and asked them to see how many matches they could make to our Public-Use File, using only those data items which we and our private detective had determined were a potential hazard. Automatch selected the most likely match for each record in the Public-Use File, and assigned a match score indicating how good the match appeared to be. Westat tested various measures of differences, finally settling on the logarithms of the differences between the items from the Master File and the items from the Public-Use File.

To our horror, the 195 records with the highest match scores were, indeed, true matches. We knew this test was based on much better data than any intruder could possibly have, but it cast doubt on the previously presumed impenetrability of the file. So, we asked Westat to evaluate which characteristics tended, more than the others, to make individual records stand out. As it turned out, the two largest factors were home mortgage interest and age, with the combination of the two being particularly damaging. After reviewing the correctly identified records, we concluded that we would have to eliminate both items from the file, or else deselect so many high-income records that all data for the highest-income classes would be questionable. So, we eliminated the fields. And because total interest deducted has only two components--home mortgage interest and investment interest--we also had to get rid of investment interest, or a user could have easily recomputed home mortgage interest. Total interest was left intact on the file.

It should be noted that, before taking this drastic step, we consulted with some of our users to see whether they would like us to preserve the interest detail on the file by introducing data-swapping between the two interest items. However, our modelers are so concerned that relationships between various fields be maintained that they would rather have us eliminate data items than get into data-swapping.

Having eliminated the cause of matching on most problem returns, we took a closer look at the remaining returns that had been identified by Automatch. In some cases, we concluded that we would have to eliminate them as outliers. Others we found could be fixed by one more round of blurring, involving only those records that had relatively high match scores and were actual matches. Having eliminated the age demographic from

the file, we also had broader classes within which we could perform the blurring.

◆ Quality of the Data

So, we finally had a file which, by very tough standards, did not appear to present a disclosure risk. But had we done major damage to the data? Obviously, there is sampling error associated with sampling the top-income classes at a rate of 1-in-3. That sampling error is easily measurable. But beyond that, was there more error? Much of the recent debate on tax reform centered on the "top 1 percent" of all income taxpayers, and the rates at which they were (and would be) paying taxes. Presumably, our users would be using our file to analyze this group. And this was precisely the group affected by blurring and elimination of outliers.

Figure 1 shows data for the top 1 percent of all taxpayers, based on size of adjusted gross income. It compares data from the Public-Use File (PUF) to those from the regular Statistics of Income file for three key items and one key ratio. It shows that the top 1 percent came out relatively unscathed by our disclosure-proofing. The most important fields to maintain, in our estimation, were adjusted gross income and tax, since much of the current debate centers around how much various proposals would affect the tax rates paid by Americans at various income levels. The differences in these estimates were well within 1 standard deviation of the value from the SOI file. The estimate of the effective tax rate changed from 28.88 percent to 28.85 percent as a result of sampling--hardly enough to cause anyone to suggest a major policy shift. The capital gains of the top 1 percent, which are taxed differently from other sources of income, also came through our process relatively unscathed. Looking at all the money amounts on the file, we found that all but 5.19 percent of our estimates for this top group were within 2 standard deviations of the SOI estimate, and all but 3.7 percent within 3 standard deviations of the SOI estimate.

We hardly had time to congratulate ourselves on our good fortune when an anonymous Harvard professor upped the ante for us: He used our 1995 Public-Use File to analyze the taxes of the top 400 taxpayers. His results, which he projected to Tax Year 2000, were published in the *Washington Post* [4]. We conferred with some of our users whom we thought might have been responsible for these estimates, and provided them with more accurate estimates from SOI's non-disclosure-proofed files--estimates that briefly got major attention in several newspapers. But the issue had been raised: Does the Public-Use File provide accurate estimates for the top 400 taxpayers?

Not unexpectedly, the effect of our disclosure-proofing on some amounts, when compared to the standard deviation, was quite a bit larger for these returns than it was for the top 1 percent. Figure 2 shows that we lost some $783 million in tax and $2.7 billion in income from the top 400. The only positive thing to be said here is that the relationship between tax and adjusted gross income--the effective tax rate--survived relatively unscathed. Overall, of the 125 money amounts shown on the PUF for the highest-income taxpayers, 38.4 percent differed from the true values by more than 2 standard deviations, and 25.6 percent differed by more than 3 standard deviations. Basically, the Public-Use File should not be used for estimates on the top 400 taxpayers.

◆ Future Plans

Our major concern at the moment, and a major topic for future research, is that we may be overdoing our disclosure-proofing by assuming that a potential intruder would have data of the quality shown on the IRS Master File. Needless to say, some of our users were unhappy when they found that the "age 65 and older" indicator was missing from the file. We would like to work with our contractor to build an accuracy indicator into the statistical matching program--in other words, an algorithm that informs the computer that data of the quality on the IRS Master File are available for only a determined percentage of the records. At the same time, we would have to continue our search for possible sources of identifiable data, and evaluate the extent to which they agree with the data on our Public-Use Files.

◆ Acknowledgments

The authors wish to thank Barry Johnson and Michael Strudler for their helpful comments.

◆ References

[1] Strudler, M.; Oh, H.; and Scheuren, F. (1986), "Protection of Taxpayers' Confidentiality With Respect to the Tax Model," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 375-381.

[2] For an excellent summary, see Scheuren, F. and Mulrow, J. (1998), "The Confidentiality Beast--A Fable About the Elephant, the Duck, and the Pig," *Proceedings of the Section on Government Statistics*, American Statistical Association, pp. 16-20.

[3] Automatch, since renamed "Integrity," is a product of Vality Technology, Inc.

[4] Kessler, G. (2001), "The Very Rich Pay Growing Tax Share," *The Washington Post*, March 15, 2001, p. E01. ■