PROTECTION OF TAXPAYER CONFIDENTIALITY WITH RESPECT TO THE TAX MODEL

Michael Strudler, H. Lock Oh and Fritz Scheuren, Internal Revenue Service

For over twenty years, the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) has made available to the public a microdata file of a sample of individual taxpayers' returns (the Tax Model). In the current climate of proposed tax law changes, this has been a valuable tool for researchers to study the effects of proposed laws and also to advance alternative proposals. The data, however, must be issued in such a form that protects the confidentiality of individual taxpayers.

A difficult problem exists in balancing protection against disclosure with providing data to the public which can give reliable analytical results. As a result of our current research on this issue, we are making several changes to our 1984 public-use file. These changes include removing certain data fields and codes from our file, altering specific codes, modifying our "blurring" process [1], and subsampling high-income returns. This paper describes the research that was involved in making these changes, and the effects these changes have on disclosure and on the statistical integrity of the data in the Tax Model file. The paper also includes a brief description and history of the Tax Model, including the importance of the disclosure issue; previous research and file changes; and recommendations for the future.

BACKGROUND

The Tax Model is a microdata file which consists of detailed information taken from a stratified sample of individual tax returns. Returns are separated into sample strata (33 in the 1983 Tax Model) based on income and presence or absence of certain schedules. Records are then selected for the file from the various strata at rates ranging from .04 percent to 100 percent. The latter strata are for high-income returns [2].

The Tax Model was first established and made available to the public in 1960. It was issued biennially through 1966, and has been issued annually thereafter. Frequent users of this public-use file include the Brookings Institution, Congressional Budget Office, National Bureau of Economic Research and the University of Michigan [3], among others. With the Tax Model to help them simulate and review the impact of tax law changes, these groups, as well as other users, have had an important impact on the national tax reform dialogue.

Because the Tax Model is issued to the public, no identifiers, such as names and social security numbers, are included on the file. Also, when state codes were added to the file (in 1978), they were limited to individuals with an adjusted gross income of less than $200,000. In the the 1960's and 1970's, these measures were considered to be sufficient for the protection of the confidentiality of individual taxpayers.

Our perception changed, however, after a reporter for the Chicago Sun-Times, purchased the 1980 Tax Model. A series of articles [4] were written in which the Tax Model was used to describe the U.S. tax system. IRS Public Affairs asked the reporter to include in his articles the caveat that the data that he used were in an unidentifiable form. He refused, stating that he, indeed, could identify some individuals. Although there is no evidence that he actually did this, under the stringent guidelines set forth for release of tax data in the Tax Reform Act of 1976 [5], SOI was challenged to research the issue.

As a result of that research [6], changes were made to the Tax Model in order to better protect the confidentiality of individuals. First, all continuous data fields were rounded to four significant digits. Second, further disguising of certain data items was deemed necessary. This disguising of data was done by a process called "masking" or "blurring" [1]. In this process, the file was independently sorted from largest to smallest value for six data items. Then for every ten records, in descending order, the average of that data item was calculated. This average then replaced the original value of that data item for each of the ten records. The "blurring" process was continued until that part of the sample with zero values for a data item was reached. These latter records were not included in the "blurring" to preserve the integrity of zero values.

The changes described above were instituted with the 1981 Tax Model and were continued in 1982 and 1983. The data fields that have been "blurred" in the Tax Model are: alimony paid, alimony received, real estate tax deductions, state income tax deductions, personal property tax deductions, general sales tax deductions, and salaries and wages ("blurred" in 1982 and 1983).

RESEARCH METHODOLOGY

Under the assumption that a data user possessed knowledge of only one data field, the test results were conclusive that the strategies taken for the 1981 Tax Model were successful in protecting taxpayer confidentiality. Indeed, at that time, even without any changes, there was virtually no chance of identifying an individual taxpayer using the data then available to the public.

In continuation of this research, the present paper examines whether knowledge of multiple pieces of information could lead to identification of taxpayers under current strategies. We started by tabulating univariate distributions on discrete values for records from the 100 percent sample strata (since this is the area which raises the most concern). We included in these tabulations some data fields on a zero-one basis and all possible code values. An example of the former is: taxpayers have royalty income (1) or they do not (0). Results of these tabulations helped indicate potential disclosure problems, generally from outlying code values. Analyzing these initial results, we selected codes and fields for tabulating bivariate distributions.

Under the assumption that the value for the selected codes and data fields are public knowledge, the bivariate tabulations indicated several disclosures might arise [7]. These results mandated that we alter our file to protect confidentiality. After analyzing the results, we are making the following changes [8] for the 1984 Tax Model:

1. Fields and codes that were selected for elimination from the files were alimony received, alimony paid, age and blindness code for primary taxpayers, and age and blindness code for secondary taxpayers. Because of the possible accessibility of accurate information on these items, these fields and codes were viewed as potential threats to the confidentiality of certain taxpayers.
2. Other codes that were considered less serious disclosure problems were altered instead of eliminated. The changes were made in an effort to balance protection of confidentiality with maintaining the viability of the file. These codes are age exemptions, marital status, and the number of children living at home. They have been altered in the following ways:

   A. Age Exemptions -- In previous Tax Model Files, this code had four possible values for every possible combination of primary and secondary taxpayer taking age exemptions. This has been changed to two values, one for the presence of at least one age exemption and one for taking no age exemptions. This change will help to prevent disclosure problems, while still allowing researchers to differentiate between returns with individual taxpayers that are 65 or over and returns with no individuals at this age.

   B. Marital Status -- In previous files, we had a separate code for widow(er) with dependent children. These returns have now been combined with joint returns with which they share the same tax rate schedule.

   C. Exemption for the Number of Children Living at Home -- Because large values for this code could cause disclosure problems, we have limited the values for this code between 0 and 3, with 3 being for all returns claiming more than two exemptions. This change still allows researchers to differentiate between "typical" families of four (married with two children) and larger ones.

3. In order to protect against a user from recalculating, with certainty, the codes that we have eliminated and changed above, we have eliminated the code for exemptions other than age or blind from the file. This has allowed us to preserve the code and field for total exemptions, which is vital to users trying to determine the effects of possible tax law changes.

TEST RESULTS

We then tested whether the combination of "blurring" plus altering or eliminating fields and codes protected the Tax Model against disclosure problems. Initially we decided to test the file using the "Spruill method" [9,10]. This is done by finding the individuals that minimize the sum of absolute deviations between the variables on a file and the actual data. This test is performed based on the assumption that the data (before changes) is known with certainty.

Before performing this test, however, we researched the field of salaries and wages to test how applicable the "Spruill method" is to our data. We selected this field because, for many prominent individuals, these data are readily accessible to the public. We used data on salaries and wages for chief executive officers published in the media. Matching salaries and wages for 93 executives listed [11,12] as earning over $500,000 with the data on our non-public use file, the average difference was over $900,000. Also, none of the cases could be matched directly, and very few could be matched with any degree of closeness with records on our file. This research indicates that using the "Spruill method" would not be a reasonable way to test our public-use file for disclosure. If we used this method, it would lead to having to adopt procedures that, for the present at least, would alter the Tax Model File far more than necessary given the outside information available.

Although researchers would have serious difficulty accurately targetting salaries and wages on our file, they might be able to construct income classes in which most of the published data would match the file. From our research and allowing for noise, we constructed the following classifications listed in Figure 1.

Figure 1.--Classification of Size of Wages and Salaries

1.	---No wages---		
2.	$1	to	$9,999
3.	$10,000	to	$199,999
4.	$200,000	to	$2,749,999
5.	$2,750,000 and over		

From previous research [6], the other field that is most accessible to the public appeared to be real estate taxes paid. These data can be found in county tax offices. This, however, is only reliable if the taxpayer owns one piece of property. Analyzing the data on our file, we established what we believed to be conservative, but reasonable classes for this field. These are listed in Figure 2.

Figure 2.--Classification of Real Estate Taxes Paid Deduction

1. $1 to $999
2. $1,000 to $1,999
3. $2,000 to $2,999
4. $3,000 to $4,999
5. $5,000 to $7,499
6. $7,500 to $9,999
7. $10,000 to $14,999
8. $15,000 to $19,999
9. $20,000 and over

We then cross-tabulated the classes for real estate taxes with the classes for salaries within subgroups of returns (depending upon various combinations of codes) to test for disclosure. The codes that we used to categorize the data were the ones we had already altered because of the potential disclosure problems (age exemptions, marital status, and number of children living at home) [13]. The seven subgroups are listed in Figure 3.

Figure 3.--Seven Subgroups of Taxpayers Based on Age, Marital Status, and Number of Children

1. Single under 65 years of age
2. Single over 65 years of age
3. Married over 65 years of age
4. Married under 65 with no children
5. Married under 65 with 1 child
6. Married under 65 with 2 children
7. Married under 65 with 3 or more children

Table 1 summarizes the seven cross-tabulations (each 9 by 5) based on the size of total salaries and real estate tax deductions for each individual high-income return (those individuals in the 100 percent strata plus all individuals with an adjusted gross income over $199,999). Approximately 5.7 percent of the cells tabulated in these tables were potential disclosure problems (cells indicating tallies of 1 or 2 individuals [7]). All of the potential disclosures were found in the highest salary classification (over $2.75 million). Although this was a great improvement over tabulations run prior to the changes we have made, this still indicates potential breaches of confidentiality. Therefore, we changed the "blurring" process as follows:

- The high-income group was separated into the seven sub-groups displayed above in Figure 3. Each group was then separated into five classes according to the salary classifications displayed above in Figure 1. (That is, the high-income individuals were separated into 35 different groups based on age, marital status, number of children and their amount of salaries and wages.) Within each of these 35 groups, the file is sorted on the "key" variable, salaries and wages. Then, the salary and wages are "blurred" one group at a time (three records at time), with no mixing of records from different groups. For example, if an individual is single (Figure 3, Subgroup 1) with a $250,000 salary (Figure 1, Class 4), this salary will only be averaged with other salaries of single taxpayers having a salary range of $200,000 to $2,749,999.

Table 1.--Size of Wages and Salaries by Size of Real Estate Tax Deductions

Cell Size	Number of Times Found	Percentage of Returns
0	14	4.5
1	8	2.5
2	10	3.2
3	5	1.6
4	8	2.5
5	7	2.2
6	6	1.9
7	6	1.9
8	3	1.0
9	5	1.6
10+	243	77.1

Source: High-Income Subsample of Individual Tax Model, 1983

- After salaries and wages are "blurred", each of the 35 groups are then sorted on a new "key" variable, real estate tax deductions. "Blurring" of real estate tax deductions (averaging three records at a time) is then done. For example, taking the same taxpayer as above (single, salary of $250,000), with a $5,500 real estate tax deduction (Figure 2, Class 5), this deduction will only be averaged with other real estate tax deductions of single taxpayers having a salary range of $200,000 to $2,749,999.

- In classifying all high-income returns into the 7 subgroups (Figure 3), we combined certain classifiers (single combined with head of household; and married over 65 with no children combined with married over 65 with children). There were too few returns to create separate subgroups for each of these.

SUBSAMPLING

Previous research by Paass [14] has illustrated that while 100 percent sampling may cause high rates of disclosures, subsamples reduce this risk considerably. Intuitively, as well as theoretically, this makes sense. In a 100 percent sample, if you target a person and find just one individual in the cell that matches information that you have on that person, you are reasonably certain of a match. However, if a subsample of a population is taken, your certainty declines in direct relation with the sampling ratio.

We elected to adopt subsampling of our 100 percent strata at a 33 percent rate. Because the lowest weighted return on the Tax Model will now be changed from 1 to 3, this rate is consistent with the "rule of 3" used for disclosure of all tabulations of tax return data [7]. Combining subsampling of high-income returns with "blurring" by subgroup, we again cross-tabulated the size of salaries by size of real estate tax deductions. We randomly selected 11 subsamples, and Table 2 displays a summary of these unweighted tabulations, using one (Subsample #6) of these subsamples.

Table 2.--Size of Salaries and Wages, by Size of Real Estate Tax Deductions, After "New Blurring" and Subsampling

Cell Size	Number of Times Found	Percentage of Returns
0	81	25.7
1	0	0
2	0	0
3	42	13.3
4	5	1.6
5	2	0.6
6	28	8.9
7	1	0.3
8	4	1.3
9	14	4.5
10	138	43.8

Source: 100 Percent Strata of Individual Tax Model, 1983

ANALYTICAL PROPERTIES

As a result of "blurring" within these 35 groups instead of all returns at one time, none of the cells (Table 2) are potential disclosure problems (at least, as we have defined data availability and disclosure). Therefore, adoption of these strategies has increased the confidentiality of individuals on our file. However, increasing disclosure protection may also cause a loss in the statistical integity of the data, as we illustrate below.

To test how the original statistical relationships of the Tax Model have been affected by disclosure protection, we compared means, covariances, and correlation coefficients for the original data, the data as "blurred" previously, and the data as changed by adopting the "blurring" strategy described in this paper.

To best assess the changes we have made, we first compared the effects of adopting different "blurring" procedures prior to subsampling our 100 percent sample (see Table 3). First, in "blurring" data, whether we use our present technique or our previous method, the method has the valuable property of being mean invariant.

Second, because "blurring" averages records together, this, in effect, eliminates extreme outliers. Adopting this strategy would then be expected to reduce the variance of a variable, and the results in Table 3 demonstrate this. Using last year's "blurring" technique, this reduction is only 13 percent for salaries and 1 percent for real estate taxes. When we invoked the new strategy of "blurring" within 35 strata, these variances decreased by 28 percent and 12 percent, respectively, for salaries and real estate taxes. Because we considered the reduction in variance for salaries to be unacceptable, we introduced the strategy of "blurring" within groups of three instead of ten. This reduced variances by only 13 percent for salaries and 6 percent for real estate taxes.

Third, in looking at correlation coefficients, researchers would probably be most interested in the effects of "blurring" on the relationships of altered data fields with income taxes. Using last year's method of "blurring," these relationships were held reasonably well (reduced by 12 percent for salaries and wages and increased by 6 percent for real estate taxes). Using 35 strata and "blurring" ten records at a time did not give as good results. However, when we reduced the number of records grouped together to three, the results were almost identical to those found last year (again, the correlation coefficient of income taxes with salaries was reduced by 12 percent and with real estate taxes was increased by 6 percent).

In comparing the new method of "blurring" (grouped 3 at a time within 35 strata) with last year's method (grouped 10 at a time within 1 strata), first and second order statistics are not considerably different. However, by also adopting subsampling of high-income returns this year, we have introduced a sampling error [15] in our estimators. To reduce this, we analyzed alternative methods to drawing a random, stratified (by the 35 subgroups) subsample. The following is a description of two of these methods, including our reasons for selecting them:

- In order to obtain deeper penetration of the stratification of returns, we randomly selected our one-third sample within zones of 12 returns. These zones were created by sorting the subsample of high-income returns on taxable income.
- Based on the assumption that the principal cause for variation between subsamples is the presence of outliers in some subsamples and their absence in others, we removed the 10 largest outliers for each of the variables analyzed (salaries, taxable income, income tax before credits, state taxes, and real estate taxes). Because of duplication, this resulted in the removal of 36 returns prior to sampling. Adopting this method would,

Table 3.--Mean, Variance, and Correlation Coefficients of Selected Variables Under Different "Blurring" Strategies

(Dollar amounts for means are in thousands and variances are in billions.)

Statistic	Salaries and Wages	Taxable Income	Income Tax Before Credits	State Tax	Real Estate Taxes
			Original Values		
Mean	$197.3	820.2	400.8	53.5	5.4
Variance	$502.7	333.4	174.4	47.0	0.6
Correlation Coefficients:					
Salaries	1.000	.273	.263	.407	.071
Taxable Income	.273	1.000	.832	.430	.137
Income Tax B/C	.263	.832	1.000	.398	.124
State Tax	.407	.430	.398	1.000	.150
Real Estate Tax	.071	.137	.124	.150	1.000
		"Blurred" Values - Last Year's Method (grouped 10 at a time within 1 strata)			
Mean	$197.3	820.2	400.8	53.5	5.4
Variance	$436.9	280.3	149.2	29.5	0.6
Correlation Coefficients:					
Salaries	1.000	.246	.241	.280	.079
Taxable Income	.246	1.000	.832	.408	.145
Income Tax B/C	.241	.832	1.000	.379	.132
State Tax	.280	.408	.379	1.000	.160
Real Estate Tax	.079	.145	.132	.160	1.000
		"Blurred" Values - Current Strategies (grouped 3 at a time within 35 strata)			
Mean	$197.3	820.2	400.8	53.5	5.4
Variance	$435.3	276.7	145.5	33.5	0.6
Correlation Coefficients:					
Salaries	1.000	.243	.235	.313	.076
Taxable Income	.243	1.000	.832	.408	.147
Income Tax B/C	.235	.832	1.000	.379	.132
State Tax	.313	.408	.379	1.000	.158
Real Estate Tax	.076	.147	.132	.158	1.000

also, offer the valuable property of increasing protection against disclosure, since outliers were shown to be our principal concern in this area. Looking at Table 4, we find that by systematically drawing our sample, standard errors are not notably improved. On the other hand, by removing outliers from the sample, the standard errors of our estimators are greatly reduced. Although removal of outliers minimizes the standard errors, it comes at the cost of losing significant data. This is illustrated by the bias [15] that is introduced when using this method (Table 4). The effect of this bias balances the reduced standard errors so that when analyzing the root mean square errors [15] of the estimators, there is little to choose between the alternatives (Table 4). Looking at the average of the relative root mean square errors of all the estimators, systematic sampling appears to be the best alternative.

CONCLUSIONS

According to our research, "blurring" by subgroups and subsampling high-income returns at a one in three rate protects the Tax Model against disclosure with a high degree of certainty, while still providing reliable data. However, we should strive for further protection of the confidentiality of individual tax returns and improvement of the statistical integity of our data. Specifically, for those outlying groups that we found to be potential disclosure problems, we should continue to research strategies to further improve our public-use file. Further research is presently being done at IRS on this issue. However, that work is in the preliminary stages and not yet ready to apply to the 1984 Tax Model.

Combining the strategies outlined in this paper (subsampling the 100 percent strata and "blurring" certain data fields within subgroups) with the measures that we had taken

Table 4.--Root Mean Square Errors and Its Components for Alternative Sampling Techniques

Variables	Root Mean Square Error	Bias	Standard Error	Relative Root Mean Square Error
	Average of 11 Samples - Stratified, Random Sampling Technique			
Averages	.0642	-.0077	.0521	.2735
Salaries and Wages:				
Taxable Income	.0681	-.0208	.0649	.2702
Income Tax B/C	.0645	-.0193	.0615	.2653
State Tax	.1535	-.1309	.0802	.5561
Real Estate Tax	.0246	.0114	.0218	.2984
Real Estate Tax:				
Taxable Income	.0374	.0283	.0244	.2267
Income Tax B/C	.0392	.0258	.0295	.2616
State Tax	.0356	.0118	.0336	.2196
State Tax:				
Taxable Income	.0838	.0138	.0826	.1887
Income Tax B/C	.0715	.0110	.0707	.1747
	Average of 11 Samples - Stratified, Systematic Sampling Technique			
Averages	.0600	-.0050	.0481	.2395
Salaries and Wages:				
Taxable Income	.0750	-.0224	.0716	.2994
Income Tax B/C	.0644	-.0109	.0635	.2558
State Tax	.1267	-.1147	.0539	.4334
Real Estate Tax	.0162	.0082	.0139	.2038
Real Estate Tax:				
Taxable Income	.0287	.0147	.0247	.1900
Income Tax B/C	.0197	.0146	.0132	.1420
State Tax	.0462	.0317	.0336	.2535
State Tax:				
Taxable Income	.0815	.0111	.0807	.1847
Income Tax B/C	.0814	.0227	.0782	.1932
	Average of 11 Samples - Stratified, Random Sample - Removal of Outliers			
Averages	.0602	-.0164	.0174	.3079
Salaries and Wages:				
Taxable Income	.0641	-.0619	.0165	.3036
Income Tax B/C	.0513	-.0477	.0190	.2387
State Tax	.2068	-.2056	.0223	1.0263
Real Estate Tax	.0348	.0322	.0132	.3371
Real Estate Tax:				
Taxable Income	.0423	.0412	.0097	.2380
Income Tax B/C	.0591	.0579	.0121	.3252
State Tax	.0319	.0286	.0140	.0140
State Tax:				
Taxable Income	.0226	.0030	.0224	.0529
Income Tax B/C	.0293	.0109	.0272	.0716

previously (rounding all data fields to four significant digits and "blurring" data), we can release the 1984 Tax Model with confidence that we are well within the guidelines of the Tax Reform Act of 1976. We are also confident that the changes that we have instituted will still retain the reliability of the data for our users' purposes.

ACKNOWLEDGMENTS

The authors would like to thank Peter Sailer for his helpful comments; Wendy Alvey and Beth Kilss, for their assistance with the oral presentation; and Nancy Dutton and Bettye Jamerson for their work on typing this paper.

NOTES AND REFERENCES

[1] Scheuren, Fritz and H. Lock Oh. "Statistical Disclosure Avoidance." Speech presented to the Washington Statistical Society on May 22, 1984.

[2] Statistics of Income Division. Individual Income Tax Returns, 1983. U.S. Department of Treasury, Internal Revenue Service, November 1985.

[3] The non-public-use version of the Tax Model is used by both the Office of Tax Analysis and the Joint Committee on Taxation for their analyses of tax-related issues.

[4] Moore, Thomas J. Chicago Sun-Times, September 26 - October 3, 1982.

[5] This act defined the type of information that could be released to the public as "data in a form which cannot be associated with or otherwise identify, directly or indirectly, a particular taxpayer", (underline added) see Haskell Amendment, Internal Revenue Code, Section 6103(b) (2)(B). Congressional Record, July 21, 1976, p. 24012.

[6] Boemio, Thomas. "Individual Tax Model Briefing." IRS Working Paper, January 1984.

[7] IRS uses the "rule of three" (meaning no cell with 1 or 2 individuals is released) to define disclosure for tabulations on the national level, see Wilson, Oliver H. and William J. Smith, Jr. "Access to Tax Records for Statistical Purposes." 1983 Proceedings of The Section on Survey Research Methods, pp.595-601.

[8] These codes and fields were changed or eliminated only for high-income returns (100 percent strata and returns with AGI over $199,999).

[9] Spruill, Nancy L. "Measures of Confidentiality." Statistics of Income and Related Administrative Record Research: 1982, Department of the Treasury, Internal Revenue Service, Statistics of Income Division, October 1982.

[10] Spruill, Nancy L. "The Confidentiality and Analytic Usefulness of Masked Business Microdata." 1983 Proceedings of The Section on Survey Research Methods, pp. 602-607, American Statistical Association.

[11] Byrne, John A. "Who Gets the Most Pay." Forbes, vol. 133, no. 13, June 4, 1984, pp. 96-146.

[12] "Executive Pay: The Top Earners." Business Week (Industrial/Technology Edition), no.2841, May 7, 1984, pp. 88-116.

[13] We also considered including the presence or absence of royalties as a potential identifier of individuals. However, after researching this field carefully, we found that you could not accurately target well-known personalities from its presence.

[14] Paass, Gerhard. "Disclosure Risk and Disclosure Avoidance for Microdata". Paper presented at the International Association of Social Sciences for Information Service and Technology, November 1985.

[15] Kmenta, Jan. Elements of Econometrics, pp. 154-171, Macmillan Publishing Co., Inc., New York, N.Y., 1971.