MEASURES OF CONFIDENTIALITY

Nancy L. Spruill, Center for Naval Analyses*

ABSTRACT

This paper proposes criteria for evaluating the minimum amount of confidentiality provided in microdata releases. They were developed for use on business data or other data for which large amounts of similar information are publicly available. The paper also uses these criteria to compare microdata releases based on five releasing strategies--adding random error, multiplying by random error, grouping, random rounding, and data swapping--using data generated from the IRS report: Statistics of Income--1977, Partnership Returns.

INTRODUCTION

IRS and other government agencies would like to release a sample of business microdata for use by researchers. However, confidentiality considerations backed by laws prohibit releasing any data that might be linked, either directly or indirectly, to an individual firm. The problem with the release of business microdata is that there are publicly available data bases that also contain business microdata and these two sources might be linked so that released data could be attributed, with high confidence, to a specific firm.

We are studying conditions for releasing IRS business microdata to researchers. We are examining releasing strategies that modify the data in ways that would leave them useful in economic studies while still satisfying the confidentiality requirements of the law. Research in this area has focused on releasing strategies where data are masked and on how such masked data can be used in analyses (Clayton & Poole [1], Rosenberg [2], Haitovsky [3], and many others). There is essentially no work on how to evaluate the confidentiality of different releasing strategies for microdata. An exception is Cox [4] who proposes how to ensure confidentiality in tabled data when a definition of breach of confidence is given. For example, if one defines a breach of confidence as tabled cells or combination of tabled cells having less than three members, as IRS does, Cox gives ways to test for violations and to eliminate these by cell suppression. He gives computational methods for implementing his work. But we want to look at how to define a breach of confidence and, in particular, a breach of confidence for microdata.

The purpose of this paper is to propose two measures of confidentiality and to use them in evaluating several releasing strategies using test data generated from the IRS report: Statistics of Income --- 1977, Partnership Returns (U.S. Dept. of Treasury [5]). The releasing strategies we will examine are adding random error, multiplying by random error, grouping, random rounding, and data swapping.

PROPOSED CRITERIA

The agency that wants to release some microdata cannot check to see how likely the data it plans to release are to match to every publicly available data base. There are too many data bases: the agency could not possibly keep track of all of them. But the agency can check to see if the released data can be linked back to the true data. This would be a conservative check to make. The agency could confidently assume that if the released data cannot be linked to the true data, then they cannot be linked to any publicly available data base. But how do we know if the released data links to the true data? We use the following strategy:

- Identify those data elements that are common to both the released data and any known publicly available data bases.
- Using the released data for one firm, compare these data with the true data for each firm, in terms of either
  - o the sum of absolute deviations, or
  - o the sum of squared differences [6]

  for all data elements identified as common to the released and publicly available data.
- Find the firm associated with the true data that minimizes this sum.
- If the firm is the same firm as the one on which the released data are based, then a link is said to be made.

Our proposed confidentiality criteria are the percent of released data firms for which a link cannot be made. High values (close to 100) indicate large amounts of confidence and small values (close to 0) indicate little confidentiality. If we want to be more conservative, we can include not only the firm associated with the true data that minimizes the sum, but also the firm that gives the second smallest value and the one that gives the third smallest value. If any of these is the same firm as the one on which the released data are based, then we say we have a link. This is the definition we use in the comparisons in the next section.

Of course, there is more to consider when choosing a releasing strategy than the amount of confidentiality it can provide. It is important to know whether the strategy can be used to give reliable analyses -- to be sure that the releasing strategy has not distorted the data beyond usefulness in subsequent analyses. But confidentiality should be

considered first because of the tax laws. Then one can select among those strategies that provide confidentiality to find those that are most useful in analyses. The analytic aspect of releasing strategies is not addressed in this paper. Spruill [7] gives a brief overview of many papers in this area. For summaries of the confidentiality issues, see the report of the ASA Ad Hoc Committee on Privacy and Confidentiality [8] and President's Commission [9] and for an extensive bibliography, see U.S. Department of Commerce [10].

Before a simple example is given, it is important to note how the criteria are defined for the "grouping" type of releasing strategy. For this strategy there can be several firms, say, 3, 5, or 10 firms, on which the released data are based. Thus if the firm associated with the true data that minimizes the sum is the same as any of these firms, we say a link has been made. And our confidentiality criteria are the percent of released data firms (average of 3, 5, or 10 firms) for which there is no link (none of the 3, 5, or 10 firms is among the minimum, 2nd minimum, or 3rd minimum).

Example

Suppose there are three firms in the population. The values of the data elements for each firm are shown in Table 1. Suppose only two data elements--net income and business receipts--are common to both the released data and to publicly available data bases. First, consider the releasing strategy that results in data for only one firm being released (strategy 1, Table 1.) Considering only the common elements, we would first compare the values for the released firm with those for Firm 1. Here the differences are (10-11) and (50-35). The sum of the absolute differences [11] is 16. And the sum of the absolute difference between Firm 2 and the released values is 8, and between Firm 3 and the released values is 11. Thus, the closest business in terms of minimum absolute deviation is Firm 2. Similarly, we look at the squared differences and find that the closest firm in terms of minimum squared deviation is Firm 2. If the data released by strategy 1 were derived from Firm 2 using some releasing strategy, we would say a correct link [12] or match had occurred and the value of our confidentiality criteria would be 0. The percent of released values where no match occurs is a measure of the amount of confidentiality in that released data.

Now suppose the releasing strategy was grouping and we released the average of the data elements for Firms 1 and 2. The released data are shown as strategy 2 in Table 1. The values of the absolute differences are 7, 7, and 2 for Firms 1, 2, and 3 respectively and the values of the squared differences are 49, 49, and 4 [13]. So Firm 3 gives the minimum value in both cases. But the data were derived from Firms 1 and 2. So no match occurs and the values of the confidentiality criteria are 100. Now suppose there were several more firms each with absolute differences greater than 7. If we then looked at the minimum sum, the 2nd minimum sum, and the 3rd minimum sum, then a link would occur and the confidentiality criteria would be zero.

OVERVIEW OF FIVE RELEASING STRATEGIES

Adding Random Error

One strategy is to add normal random error where the variance of the error (ϵ) is 1/10, 1/2, one, or two times the variance of the underlying X variable. It would seem logical to merely replace X with $X + \epsilon$. But because some variables in the data were constructed to be zero with nonzero probability, there is a less distorting strategy: Replace any nonzero data element X with either $X + \epsilon$ or 0 and any zero data element with 0 or $X^* + \epsilon$, where X^* is a random variable selected from the true underlying X distribution. The probabilities of going from nonzero to zero and from zero to nonzero are small and are chosen to keep the average of the random variable equal to the average of the uncontaminated.

Multiplying by Random Error

Similarly, for multiplying by random error, we use a non-normal error (ϵ) distribution proposed by Clayton and Poole [1], which multiplies each data element by a value between 0 and T ($T=(\alpha+2)/(\alpha+1), \alpha>-1$, α is a parameter that can be varied to give more or less protection). In particular, we replace each nonzero data element X by either $X\epsilon$ or 0 and any zero data element by 0 or $X^*\epsilon$, where again X^* is a selected random variable, and the probabilities of nonzero to zero and zero to nonzero keep the average unchanged.

Grouping

For the grouping strategy, we choose groups of size M. We use data on N x M businesses to give N average businesses. We choose one important variable--here, business receipts--as

TABLE 1: EXAMPLE

Data Element	"True" Data			Released Masked Data	
	Firm 1	Firm 2	Firm 3	Strategy 1	Strategy 2
Net Income	10.0	14.0	11.0	11.0	12.0
Business Receipts	50.0	40.0	46.0	35.0	45.0
Depreciation	5.0	6.0	6.0	5.5	5.5
Taxes Paid	0.5	1.0	0.5	0.9	0.75

the variable on which to order the sample businesses before grouping. Each released value is the average for the firms in the group except for those variables that can be zero with nonzero probability. In this case, if 60% or more businesses have zeros, the sample firm is given a zero. Otherwise, the sample business is given a nonzero value that will not change the average of the variable across all released firms. Of course, if the groups are large enough, we might be able to release the percent of the group that are zero and the average of the nonzero values.

Random Rounding

Random rounding is similar to regular rounding in that only certain values are given (e.g. integers, multiples of 1000, etc.). The difference is that the true value is not replaced with the closest rounded value but with the closest larger rounded value with probability p and the closest smaller rounded value with probability 1-p. In our comparisons we use p = .5. True values of zero are not rounded with probability .9 and follow the random rounding rules with probability .1. True values in the interval of rounded values that contains zero, are random rounded with probability .9 and are given a zero with probability .1.

Data Swapping

Data swapping occurs when certain data elements are exchanged between firms. Because we only release a subset of firms, data swapping consists of constructing composite firms using variables from several firms. In our comparisons, we use three firms to construct our released firm. We begin by selecting a subset of N firms and use the first 1/3 of the variables of each of these N firms to give the first 1/3 of the variables for our released firms. Next we search among all firms to find two firms that match each of the N firms in terms of having similar values for three key variables. We use the middle 1/3 of the variables from the first of these matches to be the middle 1/3 of the variables for our released firms and the last 1/3 of the variables from the second of these matches to be the last 1/3 of the variables for our released firms.

COMPARISON USING STATISTICS OF INCOME DATA

In this section data generated using Monte Carlo techniques and summary statistics from the Statistics of Income (SOI) publication for partnerships are used to compare releasing strategies employing the two confidentiality criteria--the sum of the absolute deviations across all common variables, and the sum of the squared deviations. The likelihood of the true business being linked to the released business is shown in the case of the most likely firm (the firm with the minimum absolute deviation or squared difference), and in the case of the three most likely firms.

Data

Thirty-six variables from the IRS report: Statistics of Income--1977, Partnership Returns [5] were selected. These variables are listed in Table 2. The first four variables are descriptive of the type of partnership and number of partners. A partnership is classified as either a limited partnership or a regular partnership. Almost 92 percent are regular partnerships. Several variables were selected to be zero with nonzero probability. For example, the SOI publication shows that about 60 percent of partnerships have zero payroll, while over 98 percent have no pension, profit sharing, annuity or bond purchase plans. Three of the variables, net income (less deficit), total receipts, and total deductions are linear combinations of the other variables.

The population of 1000 firms were constructed using the means and coefficients of variation for the 31 variables that are not indicator variables (type of partnership) or linear combinations. A set of 36 variables were constructed for each firm. A population was constructed where the variables were normal and positive or negative correlations [15] were introduced for six pairs of variables. But the population is not truly normal since some portion of the values are overridden to take account of zeros.

A 1 percent sample (ten firms) of the population is released. Our results are based on Monte Carlo constructions of 2 populations and Monte Carlo realizations of 10 released samples for each population.

Results

Table 3 gives the results. First the table is described, then the results are summarized. The table gives results for each of the five releasing strategies. The first column tells which of the two confidentiality criteria is being used -- the one based on the absolute value of the differences or the one based on the squared value. The second column simply serves to remind the reader that, in subsequent columns, the values to the left of the slash are confidentiality criteria for the one-firm, minimum value case and those to the right of the slash are criteria for the three-firm, three smallest value case. The subsequent columns show the number of common variables. They are the amount of overlap we assume between the variables we release and those in public data bases. We consider little commonality to be only 1, 2, or 3 variables; moderate overlap to be 6 or 12 variables (1/6 or 1/3 of the total number released); and, finally, total commonality of all but the type of partnership and number of partners variables.

Our results show that almost any releasing strategy provides confidentiality when only one common variable is released. But even with two or three common variables when the data are only slightly masked (adding random error, $\sigma = .1\ \sigma_x$), 1/2 or more of the firms can be correctly linked if the user has a good economic data base. More heavily masking the

TABLE 2

TEST DATA

(FROM STATISTICS OF INCOME - 1977, PARTNERSHIP RETURNS)

ITEM	PERCENT ZERO	ITEM	PERCENT ZERO
Number of Total Partnerships		Total Deductions	
Number of Limited Partnerships	91.7	Depreciation	
Number of Partners, Total		Taxes Paid Deduction	
Number of Partners, Limited	91.7	Interest Paid	
Payroll	60.4	Payment to Partners	
Net Income (Less Deficit)		Salaries and Wages	
Net Income	38.3	Rent Paid	
Deficit		Bad Debts	
Total Receipts		Repairs	
Business Receipts	6.7	Amortization	
Income from Other Partnerships		Depletion	
Nonqualifying Dividends		Cost of Sales & Operations - Total	
Interest Received		Pension, Profit Sharing, Annuity, and Bond Purchase Plans	98.3
Rents Received		Employee Benefit Programs	
Royalties		Net Loss From Other Partnerships	
Farm Net Profit		Farm Net Loss	
Net Gain, Noncapital Assets		Net Loss, Noncapital Assets	
Other Receipts		Other Deductions	

data provides more confidentiality. Both Adding Random Error ($\sigma = {}^{\sigma}x$) and Grouping (5 per group) seem to provide good amounts of confidentiality, but how much good is it for researchers to have data where the amount of error added equals the amount of error in the underlying data?

Both random rounding and data swapping seem to provide little confidentiality. This may be because of the way we are defining these strategies. And for data swapping when 1/3 of the total number of variables (12 variables) are for one firm and a large number of these are common items, a match with the true data is quite likely.

Future Plans

As we see from the results, most confidentiality problems occur when there are 9 or more common variables. A suggested way around this problem is to release subsets of variables to address specific issues. This is a good idea but requires either separate samples for each subset or few overlapping variables. Otherwise, the files might be linked and hence provide disclosure.

And, of course, there is more to a releasing strategy than the amount of confidentiality it can provide. We plan to look at releasing strategies that provide about the same amount of confidentiality and perform analyses (regression analysis, etc.) on these data using standard techniques or, where possible, the techniques developed by researchers to analyze the contaminated data.

And finally we are going to try out the releasing strategies on real tax data. We are doing that now at IRS. We are using the returns sampled as the basis of the IRS Report: Statistics of Income -- 1979 Partnership Returns [14]. This sample is relatively small, only 50,105 returns. We will be dealing with samples of this sample. But our results will tell us a great deal about the feasibility of using some of the techniques detailed in this paper.

TABLE 3.

CONFIDENTIALITY CRITERIA
MONTE CARLO FINDINGS: NORMAL-BASED DATA

Link Criteria	Number of Matches	Confidentiality Criteria: Number of Common Variables						
		1	2	3	6	9	12	32
Part 1.--Adding Random Error								
$\sigma = .1\,\sigma_x$								
Absolute Value	1/3	94/80	62/31	73/50	42/29	26/18	04/02	0/0
Square Value	1/3	94/80	63/36	76/62	49/38	56/37	27/17	06/06
$\sigma = .5\sigma_x$								
Absolute Value	1/3	96/87	93/80	96/93	85/70	73/60	46/29	02/01
Square Value	1/3	96/87	93/83	96/95	91/81	82/78	72/55	05/04
$\sigma = 1\sigma_x$								
Absolute Value	1/3	98/91	96/91	99/94	94/88	94/86	87/76	34/77
Square Value	1/3	98/91	98/93	100/99	97/95	95/92	94/86	52/38
$\sigma = 2\sigma_x$								
Absolute Value	1/3	100/96	99/97	100/97	100/98	100/98	98/96	92/80
Square Value	1/3	100/96	99/98	99/99	100/99	99/99	99/97	89/81
Part 2.--Multiplying by Random Error								
α=0, T=2.*								
Absolute Value	1/3	100/99	98/93	95/91	90/86	76/61	61/42	07/02
Square Value	1/3	100/99	97/93	95/92	90/78	78/64	62/42	09/05
α = -.75, T=5.*								
Absolute Value	1/3	100/100	99/97	100/98	98/96	99/94	97/92	82/68
Square Value	1/3	100/100	99/98	100/98	100/98	100/96	95/93	80/62
Part 3.--Grouping								
5 per group								
Absolute Value	1/3	100/99	89/73	90/83	90/88	79/70	79/63	61/40
Square Value	1/3	100/99	89/79	90/81	89/81	81/75	83/73	64/48
Part 4.--Random Rounding								
40 intervals								
Absolute Value	1/3	98/91	77/63	73/63	26/15	08/03	01/0	0/0
Square Value	1/3	98/91	76/64	75/65	39/28	16/13	10/05	0/0
Part 5.--Data Swapping								
Composite of 3 firms								
Absolute Value	1/3	0/0	0/0	68/57	18/10	01/0	21/11	07/02
Square Value	1/3	0/0	0/0	77/63	50/38	27/14	43/23	59/41

* α , T parameters of distribution proposed by Clayton and Poole (1976). (α must be greater than -1 and Clayton and Poole report that as α increases, the amount of error introduced decreases.)

ACKNOWLEDGMENTS

The author wants to thank Dave Hirschberg of SBA for his help in conducting this research. A special thanks goes to Beth Kilss and Wendy Alvey for their help in developing the ASA presentation and this paper.

NOTES AND REFERENCES

* This research was supported in part by a Small Business Association Grant #SB-1A-00075-01-1 to the Public Research Institute, a Division of the Center for Naval Analyses.

[1] Clayton, C.A. and Poole, W.K. "Use of Randomized Response Techniques in Maintaining Confidentiality of Data." Draft Report RTI Project No. 2520-1159, Research Triangle Institute, Research Triangle Park, N.C., July 7, 1976.

[2] Rosenberg, Martin Jay. Multivariate Analysis by a Randomized Response Technique for Statistical Disclosure Control." Unpublished Ph.D. dissertation, the University of Michigan, 1979.

[3] Haitovsky, Yoel. Regression Estimation from Grouped Observations, Griffin, London, 1973.

[4] Cox, Lawrence H. "Suppression Methodology and Statistical Disclosure Control." Journal of the American Statistical Association, Vol. 75, No. 370, pp. 377-385, June 1980.

[5] U.S. Department of the Treasury, Internal Revenue Service. Statistics of Income--1977, Partnership Returns. Publication 369 (4-81), U.S. Government Printing Office.

[6] In either case, normalized to mean zero and variance one.

[7] Spruill, Nancy L., "Statistical Techniques for Preserving and Evaluating the Confidentiality of Data Reseases: Literature Review", (PRI) 82-17, Public Research Institute, Center for Naval Analyses, Alexandria, VA, June 1982.

[8] ASA. "Report of Ad Hoc Committee on Privacy and Confidentiality." The American Statistician, Vol. 31, No. 2, May 1977.

[9] U.S. Government Printing Office. Federal Statistics, Report of the President's Commission, Vol. 1, 1971.

[10] U.S. Department of Commerce, Bureau of Census. "Information Privacy and Statistics, A Topical Bibliography." Working Paper 41 by Tore Dalenius, issued July 1978.

[11] For this simple example, we do not standardize the data values.

[12] For this simple example, we only look at that firm that minimizes the sum.

[13] Ties are unlikely with more than two groups and more continuous data.

[14] U.S. Department of the Treasury, Internal Revenue Service. Statistics of Income--1979, Partnership Returns. Publication 79 (3-82), U.S. Government Printing Office.

[15] Because there is no correlation information in the SOI publication, we introduced the correlations based on our intuitive understanding of the relationship between the variables.