THE CONFIDENTIALITY AND ANALYTIC USEFULNESS OF MASKED BUSINESS MICRODATA

Nancy L. Spruill, Office of the Secretary of Defense*

INTRODUCTION

In this paper, I discuss the problem of providing microdata on businesses to researchers. These microdata must satisfy two conditions: First, they must provide confidentiality to individual firms. That is, we should not be able to identify the XYZ corporation by looking at the microdata alone or in connection with other data. Second, the microdata must give reliable economic analyses. That is, if we calculate summary statistics or run regressions, we should get similar answers from the microdata as we would get using the entire data base.

At last year's ASA meetings I talked about this problem (Spruill {1}). In this paper, I'll quickly review the issues involved, the confidentiality criteria developed, and early work using test data. Then I'll present the new work using actual tax data, discuss some recommendations and conclusions, and talk about where we go from here.

ISSUES

The Small Business Administration has been developing a small business data base for policy analyses. They need microdata for individual firms to give flexibility in answering today's questions and those that will arise in the future. They gave the Public Research Institute a grant to help them expand their data base to include information on taxes. The problem was to use IRS business tax data to provide information on taxes paid, depreciation, etc. The solution was to devise releasing techniques to mask the data so that they would satisfy the confidentiality requirements of the law and would be useful in economic analyses.

The releasing technique we used was to take a subset of the data and apply a masking technique to that subset. The masking techniques included adding normal random error, multiplying by random error, grouping, random rounding, and data swapping. We modified all masking techniques to preserve zeros, which are important to researchers. As we applied the masking techniques, either 90% or 100% of the time the zeroes were unchanged. When they were changed, they were replaced by a value selected from the distribution of non-zero values for the variable. For the grouping technique, we released a zero value for a variable when 60% or more of the firms in the group had a zero value for that variable.

The economic analyses we examined were summary statistics (means, standard deviations, and percent zero), testing hypotheses about the correlation coefficient, and multiple regression analyses. Many researchers, including Clayton and Poole {2}, have looked at how "masked" data perform in economic analyses.

The problem of providing confidentiality is harder for data on businesses than for data of individuals because of publicly available data bases, such as those providing credit information on businesses. The size of the effect of these public data bases depends on (1) how many variables in the data base are "the same" as variables in the microdata we plan to release and (2) how much alike "the same" variables are. For example, how does net income used in the public data base compare to net income used to fill out the IRS form?

MEASURES OF CONFIDENTIALITY

To get a measure of how much confidentiality is provided by any releasing technique, we defined confidentiality criteria as follows:

- Select a firm {3} and mask the data so it can be released.
- Assuming the data in the public files are identical to the true data {4}, find the firm in the unmasked data that minimizes the sum of absolute deviations or squared deviations for all common variables.
- If the firm that minimizes the sum is the same as the firm {3} on which the release data are based, say a link is made {7}.
- Define the confidentiality criteria as the percent of released firms for which a link cannot be made.

(Spruill {1} calculates the confidentiality criteria using a simple example.)

In calculating the confidentiality criteria, we adjust all variables to have mean zero and variance one before the absolute deviations or squared differences are calculated. Thus, in these criteria, each variable is given equal weight. There are other confidentiality criteria that we might have used, such as one that gives different weights to different variables, or one that takes explicit consideration of zeroes and non-zeroes and only looks among firms that have the same zero pattern.

The confidentiality criteria estimate the conditional probability that we cannot use public data to identify the firm whose masked data were re-

leased, given data for the firm were released. But the chance of the data for a firm being released is small. Therefore, to get an estimate of the probability of protection of identity for any firm, we need to take into account the probability a firm is in the sample. In most cases, the sampling fraction will be between 1/100 and 1/10.

RESULTS -- TEST DATA

In order to get preliminary results and to test the computer programs, we constructed test data from summary statistics available in IRS publications {5} and looked at results for these data.

The test data consisted of 36 variables (32 economic, and 4 indicator) for a population of 1000 firms. The data were constructed using normal deviates applied to means and coefficients of variation from the IRS publication. Of the 32 economic variables we constructed 4 to be zero a certain percentage of the time and 6 pairs of variables to have non-zero correlations.

Tables 1 and 2 summarize results for the normal-based data. Note how almost all releasing strategies provide confidentiality for the case when there are only four to six common variables. However, for large numbers of common variables, only grouping provides confidentiality -- between 50 and 80 percent of the released firms do not match back to any of the five firms that were used in forming them. The estimate of the protection probability for any firm is uniformly high. It is based on a sampling fraction of 1/20 in all cases except grouping, where it is 3/20. Hence, the probability has a minimum of 95% for all but grouping, which is %85.

Table 2 shows that all releasing strategies provide good estimates of the means and percent zero, while adding random error and multiplying by random error increase the standard deviation, and grouping decreases it. Because we can keep the groups small, it does not decrease the standard deviation too much. Looking at the correlation coefficient, we see that adding random error and multiplying by random error reduces the correlation, while grouping somewhat enhances it. However, grouping and random rounding have little effect on the correlation coefficient in the cases we examined. Data swapping seems to destroy many correlations. Finally, looking at regression analyses, we find that adding random error, grouping, random rounding, and data swapping give close estimates -- adding random error slightly reduces the coefficients, but still shows significance while grouping slightly increases the coefficient and remains significant. Multiplying by random error destroys relationships for the cases we examined.

One important point needs to be made about these results. The summary statistics and some other of the economic analyses using masked data can be modified based on the knowledge of the error we introduced. So we can overcome many of the problems shown in Table 2. For example, the standard deviation is too large for the masking technique of adding random error. We know that the masking technique added normal random error with mean zero and standard deviation equal to some fraction of that in the underlying population. Therefore, we can re-estimate the standard deviation and lessen the effects of masking on the estimate of the standard deviation. Clayton and Poole {2} have done a lot of this type of work for the masking technique of multiplying by random error and Cramer {6} has looked at the effects on the correlation coefficient of the grouping technique. But what Table 2 reminds us is that for certain analyses we can't just take the data -- which looks just like real, firm-specific data -- and do analyses. For some analyses we need to modify our finding to take account of the masking.

RESULTS -- TAX DATA

We also conducted tests using actual tax return data for the tax year 1979, concentrating on partnerships in the finance, real estate, and service industries.[9] We looked at 27 economic variables for each firm. For each analysis we considered only firms within a given industry and whose business receipts were within a certain range. In order to keep the computer time small for our analyses, we limited the number of firms in our population to 1000.

We found several differences between the test data we constructed from summary statistics and that from the tax returns. First, the variances of the true variables were larger than for those we had created. Our construction of variables was based on information in the IRS publication about the coefficient of variation (ratio of standard deviation to mean). But the published coefficients of variation are for means of data for firms in narrowly defined cells of tabled data. The population of firms we used from the actual tax data are much less homogeneous, e.g., all finance firms that have business receipts less than half a million dollars. For some variables, the standard deviation of the population was more than ten times the mean. This may be

unique to the industries we examined. However, the large variation relative to the mean caused problems when it came to evaluating the masking techniques of adding random error, and random rounding. Because of the higher variances, adding one percent error equated to adding more than ten percent error in the test data.

Another difference was that many of the variables were zero a large proportion of the time. For example, for the finance industry, ten of the twenty-seven economic variables were zero over ninety percent of the time, while only five were zero less than ten percent of the time. What was really happening was that the data were bimodal or trimodal with many zero values and a few positive and/or negative ones. As a result, if too much random error was added, the masked firms had values between the modes. While that might insure confidence, the resulting data provide little information about real businesses.

Tables 3 and 4 summarize the results for actual tax data. Note how the amount of confidentiality and protection that the releasing strategies provide is much less a function of the number of common variables in the real data compared with results in the test data. This is because of the high frequency of zero values for so many variables, which reduces the probability of a correct link. For example, suppose a common variable such as net income is zero for 80% of the firms. Suppose the released firm also has a zero for net income. If we look among the firms in the population for the one with minimum difference between its net income and net income of the released firm, we'll find 80% of the firms satisfying this condition. If this were the only common variable, on average, we would be unable to link up to the correct firm at least 80% of the time. The case of several common variables could be similar if many of the common variables were zero. This is the case in our analyses, especially for large numbers of common variables. Although there are many ways to choose subsets of 27 variables, we used the following strategy to select the common variables (those in both the public files and the data we release). We ordered the variables according to how likely we thought they were to be in the public files. Variables such as ordinary income and total deductions were at the top of the list while jobs credit and net gain or loss (Form 4797) were at the bottom. In general, this ordering corresponds to an ordering on the percent zero. We then selected the first variable from the list as the case of one common variable, the first two on the list as the case of two common variables, etc. Hence, the later common variables we used are those that are frequently zero. This accounts for why the amount of confidentiality for real data does not fall off as rapidly with increasing numbers of common variables as it does for the test data.

Note in Table 4 that the results of using the masked data in economic analyses are quite similar for both test and tax data, especially for the summary statistics. However, the effect of masking on the correlation coefficient and in the regression analyses depends on p, the portion of zeroes changed to non-zeroes. Changing a zero may create an unrealistic observation. If this happens for a large fraction of the firms, it can affect the usefulness of the data in correlation and regression analyses. For example, if n = 50 and the variable is zero for forty firms (80%), on average, we will change 4 zero values to non-zero values. Only a small fraction of these firms are changed. However, to keep the mean the same, we must change 4 non-zero values to zero. This changes 4 of the 10 non-zero values, or 40%, a substantial portion of the non-zero firms. One can see that if p and the percent of zeroes are both large, all non-zero values may have to be changed to zeroes. We found that p = .1 gave little more confidentiality protection than p = 0, but it had a sizeable effect on the accuracy of the economic analyses for the tax data.

There was great variation in the results for the regression analyses even among different samples of the unmasked data. But in general we found that adding random error and multiplying by random error affected the significance of coefficients in various ways -- sometimes destroying significance, sometimes introducing it spuriously, and sometimes not affecting it. On the other hand, results for the grouping strategy show that this type of masking either destroyed significance or left it unchanged. For the actual tax data, the random rounding technique had the least effect on the regression results.

CONCLUSIONS AND RECOMMENDATIONS

Releasing data for business firms without identifiers still cannot guarantee confidentiality because of large amounts of information specific to individual firms that is available in public files. Selecting a releasing strategy requires making tradeoffs between confidentiality and accuracy in economic analyses. The most important factor needed to assess the tradeoffs is the number of variables that are common both to the released data and to public files that contain data for specified firms. But, when a large

portion of the data is mostly zeroes, the effect of common variables is reduced.

If there are few common variables, then any releasing strategy that introduces only a small amount of error to the data can be used. The strategy will produce data that will provide confidentiality to individual business firms and be useful in economic analyses of the behavior of these firms.

If, however, there are more than a few (4-6) common variables, then more care must be taken. When the data are normal-based and have small variation and few zeroes, the researcher must introduce large amounts of error to provide confidentiality -- almost as much error as the variation in the underlying data. This is true for adding random error, multiplying by random error and random rounding. However, putting data into small groups provides more confidentiality, even when the number of common variables is large. Grouping does not distort the economic analyses except for (1) the correlation between two variables, both of which are uncorrelated with the variable used to order the data before they are grouped, and (2) regression analyses where the dependent variable is uncorrelated with the grouping variable. The adverse effects of grouping on economic analyses can be reduced by carefully picking the grouping variable or using several variables (in sequence or at the same time).

When the data are less homogeneous -- with larger variation and many variables zero a large portion of the time -- the number of common variables is less important. Here, a much broader class of releasing strategies still provide confidentiality to many firms. But one must bear in mind that the effects of sampling and masking are less uniform on economic analyses. Multiplying by random error has great appeal because of the work by researchers such as Clayton and Poole on the information one can get about the underlying distribution using the masked data. Also, multiplying by random error does not depend directly on the amount of underlying variation in the data and, hence, is less influenced by outliers.

WHAT NEXT?

Our analysis shows that agencies can release masked data to researchers that will protect the identity of individual firms while providing insight into the firm's behavior through economic analyses. However, there is much more work to be done on refining masking techniques and developing statistical properties of analyses with masked data.

To refine the masking techniques, researchers familiar with the actual data should try several promising releasing techniques, such as random rounding and grouping, on tax or census data. They might try to (1) consider other variants of masking techniques, such as random rounding with 20 or 30 intervals instead of the 10 and 40 used in our analyses, (2) refine the handling of zero values, (3) define smaller populations to reduce problems of associating such diverse firms, and (4) see how "unique" firms fare in terms of confidentiality. The issue of "unique" firms is important. Researchers need to know whether "unique" firms, such as those with distinct combinations of zero and non-zero values, are given as much protection from identity disclosure as more "average" firms. If not, how are the economic analyses affected without these firms? What more drastic techniques might be used on them?

To develop statistical properties of analyses with masked data, researchers should follow the work of Clayton and Poole, Cramer, and many others. See Spruill {8} for an overview of the literature on masking techniques and the effects of masking on analysis. Almost all bad effects of masking can be substantially reduced. Some simple remedy, such as taking a larger sample, may negate many of the undesirable effects of masking.

Because the choice of which releasing strategy gives the best combination of protection and insight depends on the structure of the data, a sample of the data should be tested by the agency to estimate the amount of confidentiality that will be given to individual firms before making a data release. Hopefully, a government agency will attempt, if only on a sample basis, to apply these techniques, check for protection, and thereby be able to release useful data to economic researchers.

ACKNOWLEDGEMENTS

The authors wants to thank David Hirschberg of the Small Business Administration for his help in conducting this research. Appreciation for help with computer work and the data goes to the staff in the IRS' Statistics of Income Division Computer Room and the Corporation Special Projects Section.

NOTES AND REFERENCES

* This research was sponsored in part by the Office of Advocacy in the the Small Business administration as a grant (#SBA-1A-0007-01-1) to the Public Research Institute, a division of the Center for Naval Analyses.

{1} 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, Oct. 1982.

{2} 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.

{3} Or, in the case of the grouping strategy, the firms.

{4} Any differences between public files and true data would increase the amount of confidentiality provided by masking.

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

{6} Cramer, J.S. "Efficient Grouping, Regression and Correlation in Engel Curve Analysis." Journal of the American Statistical Association, Vol. 59, pp. 233-250. 1964.

{7} Our confidentiality criteria are more strict for the grouping strategy than for the other masking techniques -- one of the firms on which the group is based makes a link.

{8} Spruill, Nancy L. "Protecting Confidentiality of Business Microdata by Masking," Report (PRI)83-21, The Public Research Institute of The Center for Naval Analyses, Alexandria, Va. Sept., 1983.

[9] Access to identifiable tax data was made under a special consultant arrangement with the Statistics of Income Division of the IRS. The basic study was done entirely on IRS computers.

Table 1.--SUMMARY OF CONFIDENTIALITY RESULTS -- TEST DATA

Releasing Strategy	Common Variables			
	4 - 6		20 - 32	
	Confidentiality Criteria	Protection Probability	Confidentiality Criteria	Protection Probability
Adding random error ($\sigma = .5\sigma_X$)	65-85	98.0-99.0	0-5	95.0-95.5
Multiplying by random error (α=0, T=2)	90-95	99.5-99.8	10-30	96.0-97.0
Grouping (5 per group)	90-95	98.5-99.0	55-80	93.0-97.0
Random rounding (10 intervals)	5-55	96.0-98.0	0-10	95.0-96.0
Data swapping (3 firms)	20-65	96.0-98.0	5-60	95.5-98.0

Table 2--SUMMARY OF ECONOMIC ANALYSES RESULTS -- TEST DATA

Releasing Strategy	Summary Statistics			Correlation	Regression
	Mean	SD	Percent zero		
Adding random error ($\sigma' = .5\sigma_X$)	OK	too large	OK	too small	Close--but coefficient reduced
Multiplying by random error (α=0, T=2)	OK	too large	OK	too small	often destroys
Grouping (5 per group)	OK	too small	OK	OK	Close--but coefficient increased
Random rounding (10 intervals)	OK	OK	OK	OK	Close
Data swapping (3 firms)	OK	OK	OK	often destroys	Close

Table 3--SUMMARY OF CONFIDENTIALITY RESULTS -- TAX DATA

Releasing Strategy	Common Variables			
	4 - 6		18 - 27	
	Confidentiality Criteria	Protection Probability	Confidentiality Criteria	Protection Probability
Adding random error ($\sigma' = .5\sigma_X$)	90-99	99.5-99.8	75-98	99.0-99.9
Multiplying by random error (α=0, T=2)	90-97	99.5-99.9	70-85	99.0
Grouping (3 per group)	75-85	96.0-98.0	65-75	95.0-96.0
Random rounding (40 intervals)	85-95	99.0-99.8	70-80	99.0

Table 4.--SUMMARY OF ECONOMIC ANALYSES RESULTS -- TAX DATA

Releasing Strategy	Summary Statistics			Correlation	Regression
	Mean	SD	Percent zero		
Adding random error ($\sigma' = .5\sigma_X$)	OK	too large	OK	too small, depends on p	sometimes destroys, sometimes creates
Multiplying by random error (α=0, T=2)	OK	usually too large	OK	usually too small, depends on p	sometimes destroys, sometimes creates
Grouping (3 per group)	OK	too small	OK	sometimes destroys	sometimes destroys
Random rounding (40 intervals)	OK	OK	OK	slightly too small, depends on p	usually OK

p = percent of zeroes changed to non-zeroes.