DISCLOSURE CONTROL METHODS IN THE PUBLIC RELEASE OF A MICRODATA FILE OF SMALL BUSINESSES

David MacNeil and Stuart Pursey, Statistics Canada
Stuart Pursey, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

Key Words: confidentiality, disclosure control, public use microdata file

1. Background

The release of data from Statistics Canada is governed by the Statistics Act and, in the case of taxation data, also by the Income Tax Act. These two acts ensure that when data is released the confidentiality of the data and identity of individuals and businesses are protected.

The Microdata Release Committee of Statistics Canada reviews the release of microdata. In March 1997 this committee approved the release of a public use microdata file (PUMF) of financial data of small businesses. This was a first for Statistics Canada since approval for the release of a *business* microdata file had never before been obtained. The difficulty has always been related to the continuous and skewed nature of financial data. It was felt that it was too easy to deduce the identification of a business from the uniqueness of its data.

Still, the demand for the release of business microdata files has remained. In 1996 Industry Canada approached Statistics Canada. They wanted access data to improve the general knowledge and understanding of the financial structure of small businesses, to help establish a more level playing field in assessing loan risk for small businesses, and to provide data users with full flexibility in conducting custom analyses of small business performance. The data of small businesses do not contain the extreme values of the complete business sector and so it was felt that Statistics Canada could attempt to develop a small business PUMF. The attempt was successful.

The purpose of this paper is to describe the disclosure control methods used to develop the PUMF, to provide information on the data quality of the PUMF, and to offer some conclusions and ideas on the methodology.

The paper follows closely Pursey (1998) with a different emphasis on the analysis of data quality. There are many issues related to disclosure control methodology and many approaches are proposed. Eurostat (1996), United States Department of Commerce (1994), and Willenborg and de Waal (1996) provide extensive overviews.

1.1 The data

The data of the PUMF originated from a probability sample of tax records drawn from the Revenue Canada (RC) files for the tax year 1993. Small businesses are unincorporated (T1) or incorporated (T2) businesses, with gross operating revenue between $25,000 and $5,000,000.

The RC files contain business identifiers (name and address of a business); two categorical variables that classify businesses into *cells* [type of industry code (4-digit Standard Industrial Classification (SIC)] and the T1/T2 category of the business); gross operating revenue (for the T1s) or gross operating revenue, depreciation, total equity, total assets, and closing inventory (for the T2s).

The sample file also includes financial data that has been captured from the taxation data forms provided to RC. They are variables such as equity, assets (current, fixed, and total), liabilities (current, long term, and total), profit/loss, revenue, and expenses (cost of goods sold; wages, salaries and benefits; occupancy costs; and financial costs). There are 24 variables for the T1 and 38 variables for the T2.

Modifications were made to the sample file. The variables that could be used to identify the name or the geographic location of a small business were removed entirely. The 4-digit SIC variable was modified so that there were at least "s" records in each cell of the sample file and at least "t" records in its corresponding population cell. This was done by re-coding the 4-digit SIC to its 3-digit SIC, and then to its 2-digit SIC and even to its 1-digit SIC, if required, to obtain the desired counts.

2. The approach and disclosure control goals

We assume that there is an intruder attempting to link response data of any respondent to the identity of the

response data of any respondent to the identity of the respondent. Disclosure control methods are designed to prevent an intruder from making the correct link. They typically follow this approach: make reasonable assumptions about the intruder's motivation, information and tools; based on these assumptions set disclosure control goals; translate the goals to mathematical rules; implement the mathematical rules within the sample file; and measure the data quality of the resulting PUMF.

It is most difficult to anticipate the capability of an intruder. Much depends on the assumptions made about the information, tools and data available to an intruder. Moore (1996) provides a useful discussion of this issue. Each potential PUMF must be examined within its own context. It was assumed, for this PUMF, that an intruder had access to the RC population file and to record linkage tools. Four disclosure control goals were set.

a) Ensure that there is a low probability that a business from the population appears on the PUMF and ensure that an intruder cannot know that a particular business is on the PUMF.

b) Ensure that each data value of each continuous variable is perturbed and that an intruder cannot undo the perturbation.

c) Ensure that there is a low probability that a PUMF record can be correctly linked to itself on the population file and ensure that an intruder cannot know that a link is correctly or incorrectly made.

d) Ensure that unique records are removed. Nothing that maintains reasonable data quality can be done to hide these records and so they are removed from the PUMF.

These goals work together to provide an overall level of protection. Thus one might be less stringent in implementing some goals and more stringent in others as a way of maintaining better data quality.

3. Developing mathematical rules and their implementation

3.1 Goal a: There is a low probability that a business from the population appears on the PUMF

In the sample file we ensured that the probability that a record appeared on the PUMF was less than r%. This required sub-sampling in cells that contained records with sampling weights of less than 100/r. In this way an intruder, with an interest in a particular business, realises that there is a low probability that the business resides on the PUMF.

3.2 Goal b: Each data value of each continuous variable is perturbed.

The implementation of this goal was achieved in three different ways. Each data value was perturbed, the three highest data values of each variable were replaced by their average, and all data were rounded to the nearest $1000.

Perturbing data values: We explored a variety of perturbation methods. The method that worked best was to perturb each datum by a random proportion of its value subject to two constraints. The proportion generated independently for each datum falls between a threshold minimum and maximum. Also within a record, the perturbations are either always positive or always negative. The process has no impact on zeros. The method keeps the *expected values* of the means and totals unchanged but it increases the variability of the data. We experimented with a variety of values for the minimum and maximum perturbation proportion, examining the impact on goal c) that is described below. Eventually we reached a point where increasing the minimum and maximum had little impact on the record linkage results from implementing goal c). That is, the probability that an intruder makes a correct link of a PUMF record to its corresponding record on the RC file remained relatively constant. Thus it was better to stop increasing the amount of perturbation and use other techniques to meet the requirements of goal c).

Replacing the three largest data values of each variable in each cell by their sample weighted average: This modification dampens the extremes of the data. This has an impact on data quality because it removes some of the natural variability of the data but it does maintain the cell's mean and total.

Rounding all data values to the nearest $1000: This process is used to remove the trailing digits that have no real information value. For small data values this rounding dominates the impact of perturbation.

3.3 Goal c: Ensure that there is a low probability that a PUMF record can be correctly linked to its corresponding record on the population file.

The implementation of this goal has two parts: calculation of the linkage rate (given the implementation of other goals, what proportion of sample records link correctly to themselves on the population file?) and modification of data to reduce the correct linkage rate to an acceptable level.

3.3.1 Calculation

Given the implementation of goal b), an exact linkage methodology easily met the requirements of this goal. But an intruder is expected to use a more sophisticated approach. We did not use a probabilistic record linkage technique. Instead we used the nearest-neighbour approach. It is a method often used in imputation where the goal is to find a record with complete data that is similar to a record without complete data. The nearest neighbour is found by searching a pool of records with complete data and finding the one with the minimum distance from the incomplete record according to some distance function based on a set of matching variables.

Matching: For both T1s and T2s the SIC code provides a categorical *exact matching* variable. It is used to link sample cells (defined by SIC and T1/T2) to their correct population cells. Since this can be done without error, we need to reach the requirements of goal c) within each cell. On the T2 RC population file there are five continuous matching variables that can be used for finding the nearest neighbour: gross operating revenue, depreciation, total equity, total assets, and closing inventory. On the T1 RC population file there is only one continuous matching variable, gross operating revenue.

For the T2s, based on experimentation, this is the way to get the *highest* proportions of correct links from the T2 PUMF to the T2 RC file using the nearest-neighbour technique. Use the rank value transformation: replace each data value, X_i, by its rank divided by the number of data values plus one: RVT = Rank $[X_i]$ / (N+1). Also to calculate ranks, combine the sample and population files into one file and then separate them again after ranking. Use, as a measure of nearest neighbour, the sum of absolute deviations.

For the T1s, we have not linked to the T1 RC population file since there is only one continuous matching variable and perturbation itself is sufficient to meet the correct linkage threshold (this was quickly verified by a link of the PUMF to the sample file).

3.3.2 More perturbation for T2s — data swapping

The goal is to be sure that in linking the sample to the population, there is less than a p% correct linkage rate. When the correct linkage rate is greater than p% in a cell more perturbation is required. The most direct way of reducing the linkage rate to below p% is to perturb, aiming directly matching variables of records with correct links. The smallest amount of perturbation that *forces* an incorrect link is to data swap with the *second-closest* neighbour. If K is the number of correct links that are required for a p% correct linkage rate and M (>K) is the number of correct links, then data swapping is required for a further M-K records. This is done by randomly selecting the M-K from the M records.

In the small business PUMF we needed to do data swapping for about one third of the T2 records. That is, the implementation of goals a) and b) was not nearly adequate enough to protect the records. In doing a data swap on the matching variables, the relationship to certain non-matching variables is lost. Thus the data of the non-matching variables are modified, using deterministic imputation rules, to preserve the relationships. This also adds perturbation to the data.

3.4 Goal d: Using a cluster analysis to identify unique records

Some records are so unusual that no amount of perturbation (that maintains data quality) protects them. These are identified using a clustering technique: they appear in clusters with very few records. In reviewing the results of this clustering we decided not to remove any records.

4. Analysis of data quality

4.1 Methodology for the analysis

The movement from sample file to the PUMF changes the original raw data and the statistics created from them. A data quality analysis may address the original raw data (micro data analysis), the statistics generated from the raw data (macro data analysis), and the impact of each stage of disclosure control. In this paper, we examine the quality of three aggregate statistics: the mean, standard deviation, and coefficient of variation. One measure of the distance between a "before statistic", S_B, and an "after statistic", S_A is:

$$Rd_S = (S_A - S_B) / S_B$$

One can analyse the Rd_S by variable, by SIC and by T1/T2. There is a difficulty in an analysis of the variables. Since the perturbations are in the same direction within a record, the perturbations are not independent within a record. Thus it is often better to reduce the 24 T1 (or 38 T2) Rd_S to a single measure such as the mean or median.

Both an unweighted and weighted analysis can be done: not using the sampling weights helps in understanding the impact of disclosure control on "numbers" (ignoring the purpose of the data). Using the sampling weights is best for understanding the impact of disclosure control on users' ability to get results "close to" those from the unmodified sample file.

4.2 Results from the macrodata quality analysis

Farr (1997) generated a database of Rd_S values for the analyses noted in section 4.1. Metzger (1997) analysed the macrodata quality of the PUMF. The analysis here is both a summary and extension of this work. These are the Rd_S quality rating categories.

Good	$\lvert Rd_S \rvert \in [\ 0\ ,\ .05\ ]$
Fair	$\lvert Rd_S \rvert \in (\ .05\ ,\ .15\ ]$
Poor	$\lvert Rd_S \rvert \in (\ .15\ ,\ .30\ ]$
Very Poor	$\lvert Rd_S \rvert \in (\ .30\ ,\ \infty\ )$

Rd_S were calculated within each SIC by T1/T2 cell for three aggregate statistics for each of the 24 (T1) and 38 (T2) variables. Each Rd_S was transformed to its absolute value. Then the median (over the variables) within each cell was found. Table 1 provides a summary of the quality ratings of these median $\lvert Rd_S \rvert$ s.

Table 1: the frequency of the median $\lvert Rd_S \rvert$ by the quality ratings for three weighted aggregate statistics.

Weighted statistic		Good	Fair	Poor	Very Poor
Mean	T1	95%	5%	0%	0%
	T2	12%	55%	26%	7%
Std dev	T1	50%	48%	2%	0%
	T2	4%	32%	46%	18%
CV	T1	64%	33%	3%	0%
	T2	3%	45%	43%	9%

Histograms 1 to 6 show the distributions, using identical scales, of the Rd_S by T1 and T2, for three statistics of the revenue variable: the mean, standard deviation, and coefficient of variation.

Both Table 1 and the histograms show that the T1s retain much more quality than the T2s. It also shows that mean is a much better quality statistic than the standard deviation and coefficient of variation.

The conclusions are not unexpected since little perturbation is required for the T1s to meet the disclosure control goals. Data swapping, used extensively for the T2, has a detrimental effect on quality. Unlike perturbation, this disclosure control technique is not easily controlled to maintain data quality.

Metzger showed that it is difficult to find industries that are of consistently high or low quality over all statistics, all variables, and both T1 and T2. There is some consistency in the quality of the variables, perhaps due to the lack of complete independence by variable in the perturbation. The "Net Operating Profit" variable is by far the lowest quality variable. This is expected since this variable is a function of all other variables and is therefore subject to perturbation from many sources.

5. Conclusions

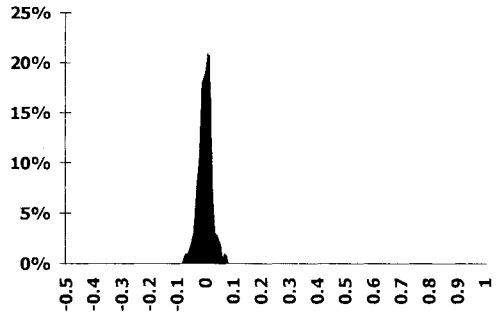
This paper has described a method of disclosure control to derive a PUMF of small businesses. The small business sector does not contain the outliers present in the complete business sector and this is what makes a PUMF feasible. The quality of the T2 small businesses is fair to poor. Most likely, the quality of a complete business sector PUMF would be of very low quality.

There is a lot of variability in the Rd_S when examined by variable, by statistic, or by industry. It would be preferable if the disclosure control process provided a more consistent and predictable impact on quality.

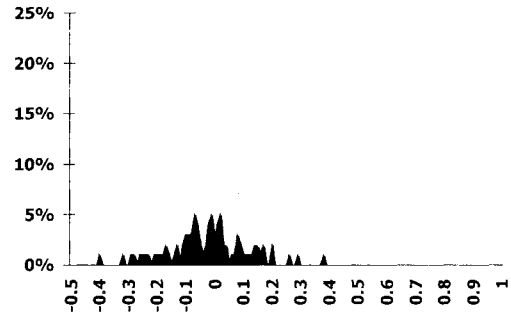
The most effective part of the methodology is the variety of disclosure control methods used to create the PUMF. This makes it extremely difficult for an intruder to untangle the perturbations.

There are several methods of linking the PUMF to the RC population file. The method that we used, nearest neighbour, is a powerful one, but others such as

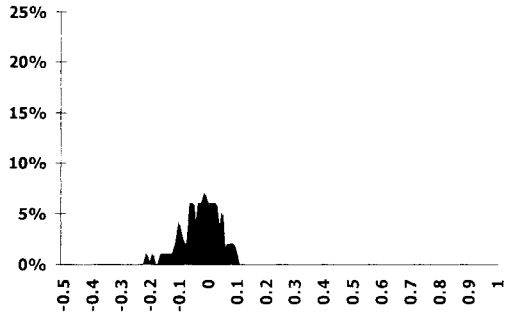
1. Distribution of the RDs for the Mean of Revenue (T1s)



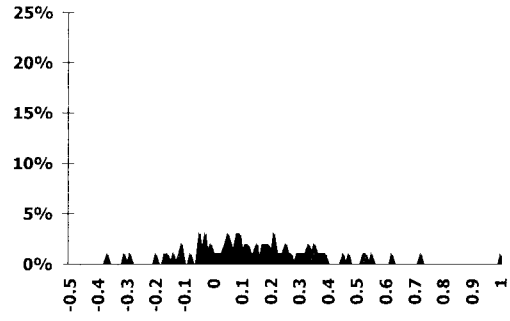
4. Distribution of the RDs for the Mean of Revenue (T2s)



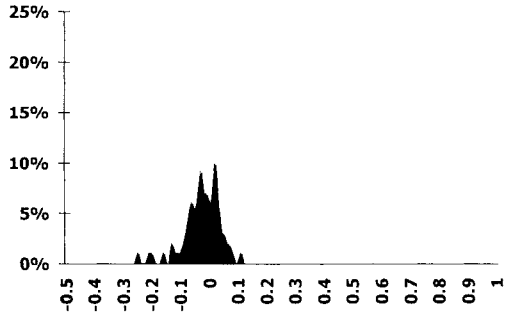
2. Distribution of the RDs for the Standard Deviation of Revenue (T1s)



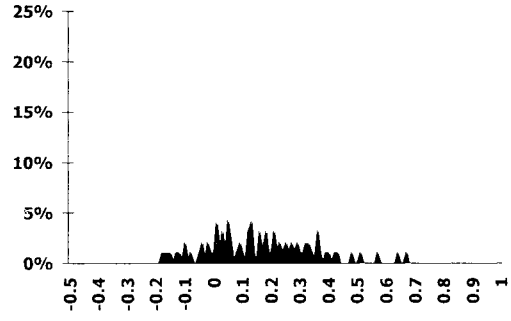
5. Distribution of the RDs for the Standard Deviation of Revenue (T2s)



3. Distribution of the RDs for the CV of Revenue (T1s)



6. Distribution of the RDs for the CV of Revenue (T2s)



probabilistic record linkage may be more powerful.

It is best to think of disclosure control as protecting the respondent from what an intruder does not already know. This is important because, as Moore (1996) argues, it is unrealistic to assume that an intruder has access to both the sample file and the PUMF.

The approach developed is a *statistical* disclosure method. Other methods of disclosure control also exist. The physical security of respondents' data is perhaps the most important and most critical of all methods. It must coexist with statistical disclosure control methods.

Further research that develops from this work should examine methods of understanding, improving, and measuring the data quality of a PUMF. In the method presented here, perhaps the sampling weights could be adjusted to calibrate the PUMF means and standard deviations to the original sample file.

References

Eurostat (1996), '*Manual on disclosure control methods*'; produced by B. Helmpecht and D. Schackis, Luxembourg: Office for Official Publications of the European Communities.

Farr, H. (1997), 'Examining Data Quality for a Proposed Method of Creating a Small Business Public Use Micro Data File', Co-operative Work Term Report, Statistics Canada and the University of Guelph.

Income Tax Act (R.S.C. 1985 5th Supplement), Canada

Metzger, R. (1997), 'Quality Analysis of a Business Microdata File', Co-operative Work Term Report, Statistics Canada and the University of Waterloo.

Moore, R. (1996), 'Analysis of the Kim-Winkler Algorithm for Masking Microdata Files—How Much Masking is Necessary and Sufficient? Conjectures for the Development of a Controllable Algorithm', Paper presented at the US Bureau of the Census - Statistics Canada Statistical Interchange, May 13-14, 1996.

The Statistics Act (1970, R.S.C. 1985, c. S19) (Canada)

Pursey, S. (1999), 'Disclosure control methods in the public release of a microdata file of small businesses', *Proceedings of the Eurostat/UN-ECE Work Session on Statistical Data Confidentiality.*

Willenborg, L and de Waal, T. (1996), *Statistical Disclosure Control in Practice*, Springer-Verlag, N.Y.