

Topic (i) - New applications of disclosure control methods

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

Submitted by Statistics Canada<sup>1</sup>

**Contributed paper**

**I. INTRODUCTION**

1. This paper discusses the disclosure control methods developed and implemented to release a public use microdata file (PUMF) of financial data of small businesses. There are five steps to create the PUMF and each is discussed in this paper: make assumptions about the capability of an intruder, set disclosure control goals based on the assumptions, translate the goals into mathematical rules, implement the rules to create the PUMF, and measure the data quality of the PUMF.

**II. BACKGROUND**

2. In March 1997, the Microdata Release Committee of Statistics Canada 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 to small businesses' 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 analysis of small business performance. The data of small businesses do not contain the extreme outliers 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 data file**

3. The original data file is a sample of tax records drawn from the 1993 Statistical Universe File (SUF), a file created by Revenue Canada, Taxation containing all taxfilers. Small businesses are unincorporated (T1) and incorporated (T2) businesses with gross operating revenue between \$25,000 and \$5,000,000. The sample file includes financial 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). The SUF includes business identifiers (name, postal code, address) in addition to the variables on the sample file. Two categorical variables are used on both files to classify records into *cells*: the industry code (4-digit Standard Industrial Classification/establishment - SICE) and the status of the business (T1 unincorporated or T2 incorporated).

4. The sample of tax file records is drawn using a two-phase sample design. The first phase is a simple random

---

<sup>1</sup> Prepared by Stuart Pursey.

sample stratified by 2-digit SICe, province, three revenue classes, and T1/T2 status. The 4-digit SICe is derived for each record in the first phase sample. The second phase is a simple random sample stratified by 4-digit SICe. The sample weight is the product of the sample weights (inverse sampling rates) from each stage. In the case of unincorporated (T1) businesses from partnerships, the weights are also multiplied by the partnership share. T1 returns associated with a partnership report for the whole business, not the individual tax filers' shares of the partnership. The partnership share provides a means of adjusting estimates for the repetition of a business in the sample file. In some cases, when the partnership share is low, the final sample weight may be less than one. Thus

$$\text{Final sample weight} = (\text{first-phase weight}) * (\text{second-phase weight}) * (\text{partnership share}).$$

### Some modifications

5. A geography variable, the postal code of the business address, is also available but it was removed entirely from the sample file. This helps in reducing disclosure risk and has almost no impact on data quality, as this variable explains only about 2% of the variability of gross operating revenue. The 4-digit SICe variable is modified so that there are at least  $s$  records in each cell of the sample file and at least  $t$  records in its corresponding population cell. This is done by recoding the 4-digit SICe to its 3-digit SICe, and then to its 2-digit SICe and even to its 1-digit SICe, if required, to obtain the desired counts.

### III. APPROACH

6. We assume that there is a person who attempts to link response data of any respondent to the identity of the respondent. That person is called an intruder. Disclosure control procedures are designed to prevent an intruder from making the correct link. Disclosure control methods 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
- measure the data quality of the resulting PUMF.

### IV. THE INTRUDER

7. It is most difficult to understand the capability of an intruder. Much depends on the assumptions made about the information, tools and data available to an intruder. Each potential PUMF must be examined within its own context.

8. *Motivation of an intruder.* There is one common motivation to all intruders. The intruder, by accident or otherwise, may attempt to identify the respondent associated with a particular data record or else attempt to find the identify of *any* respondent on the file. Thus the goal of a disclosure control method is to prevent correct links of a respondent's data to the respondent's identification.

9. *Information available to an intruder.* Following the discussion in Moore (1996), one might assume that the intruder has all of the following: access to the sample file before the application of disclosure control methods, an understanding of the types and occurrences of nonsampling errors in the sample and the edit and imputation rules used to deal with them, knowledge of the disclosure control methods used to create the PUMF, access to a SUF where business identifiers reside, and knowledge that a link of a PUMF record to the SUF is, in fact, the correct link. Moore feels that these assumptions are not realistic. He also notes that Muller, Blien and Wirth (1995) share his opinions.

10. Clearly it is not realistic to assume that the intruder has access to the unperturbed sample: if an intruder has access there is no use and no need for a PUMF. It is most unlikely that the intruder has knowledge of the nonsampling errors and the edit and imputation rules applied to the sample file. Statistics Canada keeps the details of the disclosure control methods confidential. Still, it is important to say that disclosure control processes have been applied to the PUMF to discourage intruders and to reassure respondents. But it is realistic to assume that the intruder has access to some population file where business identifiers reside with several matching variables as some government departments have access to the SUF or a variant of it.

11. *Tools of an intruder:* We assume that the intruder has access to record linkage software.

## V. DISCLOSURE CONTROL GOALS

12. Four disclosure control goals were set:

- a) Ensure that there is a low probability that a business from the population appears on the PUMF. The purpose is to cast doubt in an intruder's mind that a particular small business even exists on the PUMF.
- b) Ensure that each data value of each continuous variable is perturbed. The purpose is to cast further doubt in an intruder's mind. If an identification link is made C correctly or incorrectly C the intruder knows that the data of the record is not the *exact* data of the respondent C it has been changed in some way.
- c) Ensure that there is a low probability that a PUMF record can be correctly linked to itself on the population file. The purpose is to make it difficult for an intruder to make correct links from the PUMF to a SUF.
- 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.

## VI. DEVELOPING MATHEMATICAL RULES AND THEIR IMPLEMENTATION

### Goal a

13. 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 subsampling in cells that contained records with sampling weights less than 100/r (excluding the impact of the partnership share). Thus the final PUMF weight for a record is

$$(\text{first-phase weight}) * (\text{second-phase weight}) * (\text{partnership share}) * (\text{subsample weight}).$$

### Goal b

14. 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.

15. *Perturbing data values:* We explored a variety of perturbation methods. The method that worked best is this: provide independent random noise as a proportion of each data value with the only limitation that the minimum and maximum proportion of random noise was constant. The process has no impact on zeros. The method keeps the *expected values* of the means and totals unchanged but increases the variability of the data. Thus a random number is selected from the interval [a,b] and then with probability 0.5 it is changed to its negative. More formally a data value is modified so that it is placed randomly on the interval

$$[ x - bx, x - ax ] \text{ or } [ x + ax, x + bx ]$$

where

- x = original data value,
- a = minimum perturbation as a proportion of x,
- b = maximum perturbation as a proportion of x.

16. We experimented with a variety of values for a) and b), examining the impact on goal c) that is described below. Eventually we reached a point where increasing a and b had little impact on the record linkage results from implementing goal c). That is, the probability of a correct link of a PUMF record to its corresponding record on the SUF remained relatively constant. Thus it was better to stop increasing the amount of random noise and instead use other techniques to meet the requirements of goal c).

17. *Averaging the three highest data values of each variable in each cell* This modification perturbs the three largest data values of each financial variable in each cell, thus dampening the extremes of the data. This is done by creating a sample weighted average of each variable in each cell. 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.

18. *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 (less than \$10,000) this rounding dominates the impact of adding random noise.

### Goal c

19. It is important to 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.

### Calculation

20. 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 any kind of probabilistic record linkage software such as Statistics Canada's Generalized Record Linkage System (GRLS) although perhaps an intruder might have access to such a software package. We used the nearest-neighbour approach in Statistics Canada's Generalized Edit and Imputation System.

21. *Matching variables*: For both T1s and T2s the SICe code provides a categorical matching variable; it is used to link sample cells (defined by SICe and T1/T2 status) to their correct population cells. On the T2 SUF there are five continuous matching variables: gross operating revenue, depreciation, total equity, total assets, and closing inventory. On the T1 SUF there is only one continuous matching variable, gross operating revenue.

22. *Matching*: Since (acting as the intruder) we are able to link a PUMF cell to its correct cell on the SUF, we calculated the proportion of correct links and applied within each cell the data swapping technique described in 5.3.2.

23. T2s: Based on experimentation, this is the way to get the *highest* proportions of correct links from the T2 PUMF to the T2 SUF.

- Use nearest-neighbour linkage
- Use the rank value transformation: replace each data value by its rank divided by the number of data values plus one:  $RVT = \text{Rank}[X_i] / (N+1)$ . Also, to calculate ranks, combine the sample and population files into one file and then separate them again after ranking.
- Other transformations (mean and standard deviation, median and inter quartile range) work almost as well; even no transformation works well.
- Use, as a measure of nearest-neighbour, the sum of absolute deviations.
- Other measures of distance work almost as well (sum of squared deviations, minimum of the maximum absolute deviation).

24. T1s: We have not linked to the T1 SUF since there is only one continuous matching variable. Actually there is no need because we matched the T1 PUMF to the T1 sample file. The linkage rate was far below the p% threshold. This happens because there is just one variable to match and random noise itself is sufficient to meet the threshold.

### More perturbation C data swapping

25. The goal is to be sure that in linking the sample to the population, there is less than a p% correct linkage rate. When, in a cell, the correct linkage rate is greater than p%, more perturbation is required. The most direct way of reducing the linkage rate below p% is to perturb, aiming directly at the data of matching variables of 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.

26. In the case of the small business PUMF we needed to do data swapping on about a third of the records. That is, the implementation of goals a) and b) was not nearly enough to protect the data records. In doing a data swap on the matching variables, the relationship to certain nonmatching variables is lost. Thus the data of the nonmatching variables are

modified (through deterministic imputation rules) to preserve the relationships. This also adds perturbation to the nonmatching variables.

### Using a cluster analysis to identify unique records

27. 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. Transforming the data before clustering may be a more effective way of finding unique records; this has not been tried.

### Summary of the disclosure control method

28. Suppose that an intruder does announce that a correct link has been made. At Statistics Canada we will be reassured that:

- the PUMF sample weight of each record in each cell is greater than or equal  $100/r$ , implying that there is a low probability ( $r\%$ ) that any particular record of the SUF exists on the PUMF,
- if the record is on the SUF, there is less than  $p\%$  chance that a correct link has been made,
- if a record is on the SUF and if a correct link has been made then the data values are in error by a proportion between  $a$  and  $b$ , data swapping has possibly added more perturbation, the three largest data values of each variable in each cell have been averaged together, all data have been rounded to the nearest \$1000 and any unique records (as determined by a cluster analysis) have been removed.

STEP	WHAT	HOW	WHY	Impact on Data Quality
1 (File 1)	Unperturbed sample file.			
2 (File 2)	No geography code.  $\geq s$ sample records and $\geq t$ population records in each cell (SICe x T1/T2 Status).	Remove geography code.  Recode SICe from 4-digit to 3, 2, or 1 digit to meet the minimum cell size requirements.	Removes geography as a factor in disclosure risk.  Helps in applying the disclosure control techniques.	Little impact as geography explains little of the variability of financial variables.  But it is not possible to produce estimates by geography.
3 (File 3)	Subsample of File 2.	Subsample so that the probability that a record appears in the PUMF is less than $r\%$ .	Adds doubt in an intruder's mind that a particular population record is on the sample file.	Since the sample size is now lower than before, statistics from the sample file are less reliable.
4 (File 4)	First version of a perturbed file.  "Random noise" is on each continuous datum.	Pick a uniform random number from $a$ to $b$ and change it to its negative with probability 0.5.  Apply that number as percentage change to the unperturbed data value.	Ensures that if an intruder says that an identification is made, then the intruder has the wrong data values by a proportion of $a$ to $b$ .	Keeps the expected values of the means and totals the same. Has minimal impact on percentiles. Increases the variability of the data - leading to slightly less reliable estimates.
5	Link sample to population.  T2 (5 matching variables)  T1 (1 matching variable)	Nearest-neighbour link using the L1 Rank Value Transformation (the minimum sum of absolute deviations based on ranks)  Each data value has been replaced by its rank divided by the number of data values plus one: $RVT = Rank [X_i] / (N+1)$  To calculate ranks the sample and population files are combined into one file and then separated again after ranking.	Need to calculate how often an intruder can correctly link a sample record to itself on the population file.  The goal is to get less than a $p\%$ linkage rate in each cell.	Comments on linkage methods:  1. Nearest-neighbour linkage methods are extremely powerful 2. We found that the transformation to rank values gave the highest correct linkage rate 3. There are several methods to measure distance in nearest-neighbour linkage:  L1: Sum of absolute deviations L2: Sum of squared deviations  L $\infty$ : Minmax

				In our case L1 gave the highest linkage rates.
6 (File 5)	Data swap on matching variables.	Use the second closest nearest-neighbour as the donor record.  Do only enough times so that within each cell at most p% of links to the population file are correct.  Data swap on the matching variables.  Nonmatching variables are modified to be consistent with the data swap variables by deterministic imputation.	This is required to ensure that the percentage of correct links from the sample to the population is below a threshold.  Thus an intruder who says that a match is made has only a p% chance of being correct.	Using the second-closest nearest neighbour link helps to preserve data quality yet still forces an incorrect link.  The method does not maintain the expected values of the totals of the variables.  Since only matching variables are swapped, the relationships between matching and nonmatching variables are perturbed.
7 (File 6)	The 3 highest data values of each variable in each cell are averaged.	Average the top 3 data values of each variable in each cell using the sampling weighted average.	This reduces the visibility of outliers, making them less extreme.	This has no impact on totals and averages but reduces the variance of the data and takes away some of the natural skewness in the data.  It lowers the highest data value and raises the third highest data value. The second highest data value may go either way.
8 (File 7)	Unique records (determined by a cluster analysis) are removed.	Use PROC FASTCLUS in SAS to identify clusters with very few records -- remove these records. PROC FASTCLUS is sensitive to the number of clusters asked for and so it was tried for 5, 10, 15 & 20 clusters.	These records, identified by a cluster analysis, are those for which no amount of perturbation (that maintains data quality) can protect their identities and so they are removed.	After reviewing the results of the clustering we decided not to remove any records at all.
9 (File 8)	All data values are rounded to the nearest \$1000.  Provide a record ID.	Round all detail variables to the nearest \$1000 and recalculate all subtotals and totals.  Randomly order the records and then assign a record ID of 1 to N.	This provides further perturbation to the data.	Little impact on data quality.
10	Link sample to the population.	See method of step 5	This provides a final check that the linkage rate is less than r% in each cell.	Not applicable.

## VII. ANALYSIS OF DATA QUALITY

### Methodology for the analysis

29. 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 the above at each stage of modification.

30. Micro-analysis: In examining the original raw data for a given variable we want to measure the distance between the “before data”,  $x_{Bi}$ , and the “after data”,  $x_{Ai}$ , for a unit  $i$ . One measure is the Relative Distance:

$$Rd_i = (x_{Ai} - x_{Bi}) / (x_{Ai} + x_{Bi})$$

This type of analysis is good for understanding what has happened to the raw data, perhaps without much interest in the purposes of the data. Given  $Rd_i$  one can create its distribution and estimate its mean, standard deviation, coefficient of variation, median, range, correlations between variables, and other descriptive statistics of interest. This can be done by SICE, T1/T2 status, variable, and size of businesses.

31. Macro-analysis: One typical use of the microdata file is to create statistics such as the mean, standard deviation, percentiles and coefficient of variation for a given variable, and correlations between variables. Again we want to measure

the distance between a “before statistic”,  $S_B$ , and an “after statistic”,  $S_A$ . The Relative Distance is:

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

One can compare the  $Rd_S$  by variable, SICE and T1/T2 status. Both unweighted and weighted analyses can be done: not using the sampling weights helps in understanding the impact of the disclosure control method on “numbers” (that is, ignoring the purpose of the data). Using the sampling weights is best for understanding the impact of the disclosure control method on users’ ability to analyse the data of the PUMF and get results “close to” those obtained from the unmodified sample file.

32. In the PUMF there are about 4000 “sets” of data, where a “set” consists of the data values of a particular variable within a particular T1/T2 and SICE cell. A method of summarizing these  $Rd_S$  data is required. One approach is to provide a

$$MSE_S = (\overline{Rd_S} - \overline{RD_S})^2 + \frac{1}{K} \sum_{k=1}^K (Rd_{S(k)} - \overline{Rd_S})^2$$

frequency distribution of the size of the  $Rd_S$ . Another is to calculate an interpretation of the Mean Square Error of the  $Rd_S$ : where  $Rd_{S(k)}$  is the value of  $Rd_S$  for “set”  $k$  ( $k=1, \dots, K$ ),  $\overline{RD_S}$  is the unknown true mean of the  $Rd_{S(k)}$  and  $\overline{Rd_S}$  is the mean of the  $Rd_{S(k)}$ .

We want the disclosure control method to be such that the unknown true mean of the  $Rd_{S(k)}$  is zero and so we set it to the

$$MSE_S = \overline{Rd_S}^2 + \frac{1}{K} \sum_{k=1}^K (Rd_{S(k)} - \overline{Rd_S})^2 = \frac{1}{K} \sum_{k=1}^K Rd_{S(k)}^2$$

desired value, zero. Thus the data quality measure is

We concentrate on the value  $\overline{Rd_S}$  in the first term of the middle expression above as it is both the expected value and the bias of the  $Rd_{S(k)}$ .

33. Each stage of the disclosure control process has an impact on data quality. Measuring data quality at each stage helps in understanding the “value” of each disclosure control element. That is, for a given set of disclosure control elements, we would want to use a combination of them so that the loss in data quality is at a minimum.

### Results from the macro data quality analysis

34. Farr (1997) developed a set of SAS programs to generate a database of  $Rd_S$  values for almost all the scenarios described above. Metzger (1997) analyzed the macro data quality of four statistics: the mean, the standard deviation and the coefficients of variation for each variable, and the correlations between the variables. A summary of his analysis is shown in Tables 1 and 2. These are the quality rating categories for an  $Rd_S$  of a “set”:

Good	:	$-0.05 \leq Rd_S \leq +0.05$
Fair	:	$-0.15 \leq Rd_S < -0.05$ or $+0.05 < Rd_S \leq +0.15$
Poor	:	$-1 < Rd_S < -0.15$ or $+0.15 < Rd_S < +1$
Very Poor	:	$Rd_S = +1$ or $-1$

35. Table 1a provides a frequency distribution of the  $Rd_S$  quality ratings of the “sets” cross tabulated by T1/T2 status and by statistic. For example, 65.9% of all “sets” involving T1 data are classified as good for the (unweighted) CV statistic.

36. Table 1b shows a count, by T1/T2 status and by statistic, of the number of variables classified to negative, zero, and positive bias. For each statistic the “set”  $Rd_S$  were calculated and averaged over all “sets” representing a common variable and T1/T2 status. The sign of the average gives an overall sense of the direction of the bias. Thus, for example, 21 of the 24 T1 variables have a negative average  $Rd_S$ , or bias, for the (unweighted) CV statistic.

*Table 1 : Results of Analyses from the Approximately 4000 "Sets"*

Statistic		<i>Table 1a</i> Percent frequencies of the Rd Data Quality Categories				<i>Table 1b</i> Bias component of the MSE: # of variables with negative, zero, and positive $Rd_s$ (bias)		
		Good	Fair	Poor	Very Poor	Negative	Zero	Positive
CV	T1	65.9	19.6	13.8	0.8	21	2	1
	T2	38.3	41.4	19.1	1.2	10	10	18
Weighted CV	T1	69.8	16.7	12.8	0.7	20	1	3
	T2	36.1	40.5	22.2	1.2	9	1	28
Mean	T1	77.3	13.4	8.5	0.8	13	11	0
	T2	46.1	34.2	18.5	1.2	36	1	1
Weighted Mean	T1	85.8	7.8	5.7	0.7	12	12	0
	T2	44.1	35.1	19.7	1.2	37	0	1
Std. Dev.	T1	59.4	23.3	16.7	0.6	22	1	1
	T2	28.1	35.9	35.7	0.4	25	2	11
Weighted Std. Dev.	T1	67.0	19.3	13.1	0.6	22	1	1
	T2	30.8	36.3	32.6	0.4	14	5	19

37. Table 2 provides counts of the quality ratings of the correlation statistic for the 24 T1 and 38 T2 variables. Each T1 variable generates 23 correlations and each T2 variable generates 37 correlations. The correlations are calculated over all data points, that is, without regard to SICE. For each T1 or T2 variable the mean of the absolute values of the 23 or 37  $Rd_s$  is assigned a quality rating category.

*Table 2* Number of variables in each data quality category for the correlation statistic

	Good	Fair	Poor	Very Poor
T1 variables	3	14	7	0
T2 variables	0	12	26	0

38. T1 statistics maintain much more data quality than the T2 statistics. This is expected since little perturbation is required for the T1s to meet the disclosure control goals. Tables 1a and 1b show that there is little difference in the quality between weighted and unweighted statistics and not too much difference between the three statistics (although the mean statistic maintains the best data quality and the standard deviation suffered the most). The bias for the mean statistic and the standard deviation statistic is often negative (that is, a sample file statistic is generally larger than a PUMF statistic). Metzger (1996) showed that it is difficult to find particular industries that are of consistently high or low quality over all statistics, all variables, and both T1 and T2. But there is some consistency in the quality of the variables. The "Net Operating Profit" variable is by far the worst quality variable. This is not unexpected since this variable is a function of all other variables and is therefore subject to perturbation from many sources. Metzger (1996) found a lot of variability in the  $Rd_s$  over the 4000 cells, when grouped by variable, by statistic, or by industry. It would be better if the disclosure control process provided a more consistent and predictable drop in quality.

## VIII. CONCLUSIONS

39. 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.

40. There are several methods of linking the PUMF to the SUF. The method that we used, nearest-neighbour, is a powerful one, but others, for example the approach used in Statistics Canada's Generalized Record Linkage System, may be

more powerful.

41. It is not difficult to implement the disclosure control process on a microcomputer using SAS but it does require intricate SAS programming, a very competent SAS programmer, and a significant amount of time available for implementation.

42. 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.

43. The approach developed is a *statistical* disclosure method. Others 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.

44. The analysis shows that the quality of T1s is good to fair and the quality of T2s is fair to poor: the T1s have retained much more quality than the T2s. This is expected since the T2s require much more perturbation than T1s to reach the disclosure control goals.

**References**

- 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.
- 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.
- Muller W., Blien, U. and Wirth, H. (1995). Identification Risks of Microdata. *Sociological Methods and Research* 24, 131-157.