DISCLOSURE CONTROL STRATEGY FOR THE RELEASE OF MICRODATA IN THE CANADIAN SURVEY OF LABOUR AND INCOME DYNAMICS

Christian Nadeau, Éric Gagnon, Michel Latouche, Statistics Canada
Christian Nadeau, Jean Talon Building, 5th Floor, Ottawa, Ontario, K1A 0T6

Key Words: Longitudinal survey, Income data, Microdata modifications, Random rounding.

1. INTRODUCTION

For more than 30 years, the Canadian Survey of Consumer Finances (SCF), a cross-sectional survey, had been responsible for producing income data on individuals and families, including low income measurements. The survey released, on a yearly basis, several publications covering various aspects of individual and family income, as well as individual and family level public use microdata files (PUMF), which were hierarchical in structure.

In 1993, Statistics Canada implemented the longitudinal Survey of Labour and Income Dynamics (SLID) in order to support studies of the economic well-being of individuals and families and of their determinants over time. In order to harmonize individual and family income statistics and to reduce production costs, it has been decided to integrate the longitudinal and the cross-sectional surveys into a unique one (Latouche *et al.,* 1997 and Webber *et al.*, 1999). In 1998, SLID became the official source of longitudinal and cross-sectional income data on individuals and families. This means that, in addition to the support of longitudinal analysis of income data, SLID should also deliver the same line of products SCF used to do, including a hierarchical set of cross-sectional PUMFs, although the sample is longitudinal.

This paper presents the strategy that is planned in order to provide external users the access to microdata on individual and family income. The emphasis is put on the release of a set of hierarchical cross-sectional microdata files containing income information from a longitudinal sample. A brief description of the survey is provided in section 2. The dissemination strategy and the different factors of disclosure risk that are particular to SLID are described in sections 3 and 4 respectively. In section 5, we introduce the disclosure control strategy used and we present some of the disclosure control methods to be applied to the income variables. In sections 6 and 7, we respectively discuss the evaluation of the effect these methods have on the risk of disclosure and on survey data analysis. 1995 and 1996 data are used in order to do so. We finally discuss the future work in section 8.

2. SURVEY DESCRIPTION

SLID is collecting data for two panels of 15,000 households each (Lavigne and Michaud, 1998). The panels are surveyed each year for six consecutive years. The first panel was introduced in 1993, and a second panel was introduced in 1996. A new panel is introduced every three years to replace the older panel, which implies a three year overlap between two consecutive panels.

People who are selected at the beginning of a panel are considered as longitudinal respondents and they are part of the panel for its six years of duration. Likewise, people who start to live with a longitudinal respondent after the selection of the panel are also included in the survey, as long as they live with a longitudinal respondent. They are referred to as the cohabitants.

As part of the survey, detailed information is collected on different income sources, labour, education and on many other personal and family characteristics. In order to collect income data, Statistics Canada offers the respondents to either report income through an interview or to give permission to link to their income tax file. This last option is used in 75% of the cases. A complete description of the content can be found in the SLID user's guide (Statistics Canada, 1997).

3. DISSEMINATION STRATEGY

Prior to the integration, the SLID dissemination strategy consisted in the release of an annual set of PUMFs. This set included a person level and a job level cross-sectional file, and a person level and a job level longitudinal file. The cross-sectional files contained variables related to the current reference year, while the longitudinal files contained variables for the first to the most up to date reference year. Both files also contained fixed variables. The cross-sectional and longitudinal files were very similar. In terms of content, the corresponding longitudinal and cross-sectional files contain almost the same variables, apart from the fact that cross-sectional files contain variables for only one reference year, contrary to longitudinal files. In terms of the individuals appearing on these files, the difference was also very small. For example, the cohabitants only appeared on the cross-sectional files. A particularity of the pre-integration

PUMFs was that it did not allow household reconstitution. It was considered too difficult to protect the confidentiality of a household level longitudinal microdata file (Lavallée and Grondin, 1994 and Grondin, 1995).

Due to the integration, it is now becoming a priority to release cross-sectional PUMFs that meet the former SCF users' requirements. In order to do so, the content of the cross-sectional PUMFs needs to be redefined. One important difference with the previous cross-sectional PUMFs is that they will include household and family identifiers, which will allow household and family reconstitution. In addition to the person level file, a household level and a family level file are also produced. The household and family identifiers will allow one to link every record on the person level file to their corresponding record on the household level and family level files.

The release of such cross-sectional PUMFs compromises the release of longitudinal PUMFs. It is still thought that the risk of disclosure associated to a longitudinal file that allows household and family reconstitution would be too high and that it might not be possible to create a longitudinal file for which the risk of linkage with the cross-sectional file is low. Such linkage, even if the longitudinal file does not include household or family identifiers, would allow family reconstitution. For priority reasons, it was decided that it is better to concentrate on the release of the cross-sectional PUMFs and to consider other options to make longitudinal data available.

Among the possible options are the remote access of data that might include the release of synthetic files, and the creation of research centres. In the remote access option, the searchers use the synthetic file to validate their computer programs, then ask Statistics Canada to run them with the real files. The outputs are scanned before they are sent back to the searchers. On the other hand, the implementation of regional research centres allows the searchers to directly access the real data as if they were Statistics Canada employees. However, the searchers must comply to the Statistics Act especially as regard secrecy.

In the rest of the paper, we will focus on the release of the cross-sectional PUMFs.

4. RISK OF DISCLOSURE

The release of microdata usually entails some risk of disclosure. Before we discuss this risk in the context of the release of SLID PUMFs, we need to adopt a definition of disclosure. Two different types of disclosure are usually distinguished in the literature: disclosure by re-identification and disclosure by attribute (de Waal and Willenborg, 1996). In this paper, we concentrate on disclosure by re-identification. We say that there is re-identification when an intruder uniquely and correctly identify a unit on a PUMF. We say that there is disclosure by re-identification when the intruders learn some confidential information on this unit. To achieve re-identification, the intruder needs to use prior knowledge he or she has about some of the units of the population, which are included in the survey sample. This knowledge might have been acquired through the access to some identification files.

In the case of SLID, three factors were identified as the most problematic ones in terms of the risk of disclosure. The first one is the household and family reconstitution possibility offered by the PUMFs. This factor results in an increase of the amount of information available on individuals, which results in an increase of the possibilities of re-identification.

The second factor is the fact that the collected information is available year after year (for up to six years) for the major part of the sample. The successful linkage of units from two consecutive files does not constitute a case of disclosure in itself, but it would lead to the creation of a "longitudinal" file, which we previously decided to avoid because it was thought to be too risky regarding re-identification.

The third factor is the use of the Income Tax Data File (ITDF) as a data collection tool. This factor results in an increase in the accuracy of the information that can be used for re-identification.

In the next section, we mainly discuss methods for controlling the risk associated with the last two factors. The emphasis is put on the income variables in the remainder of this paper.

5. DISCLOSURE CONTROL

Before the identification and application of disclosure control methods, the variables are partitioned into three sets: the direct identifiers, the indirect identifiers and the sensitive variables (de Waal and Willenborg, 1996). The direct identifiers, such as names, addresses and design variables, are dropped from the PUMFs. The disclosure control methods are then applied to the indirect identifiers to reduce the risk of re-identification, and to sensitive

variables to reduce the risk of disclosure associated with re-identification. In the case of SLID, we have decided to include the income variables in both the set of indirect identifiers and the set of sensitive variables.

Disclosure control methods are usually classified into two categories: data reduction methods and data modification methods.

Among the data reduction methods used, some indirect identifiers are dropped from the PUMFs at the moment of the content determination. For example, the province of residence is the only level of geographic information that appears on the PUMFs. Other variables that are not essential to cross-sectional analysis are also dropped from the file. In addition to variable suppressions, category grouping is also performed for some variables. This is the case for variables such as industry and occupational codes. Grouping is also done for numeric variables such as number of jobs. In such a case, only the value 0, 1 and 2 are possible. Any individual with more than two jobs is coded to 2.

Data modification methods that are used for the creation of SLID cross-sectional PUMFs mainly address the case of numeric variables, namely the income variables and the year of birth and age variables. In the case of the year of birth and age variables, noise is randomly added to the reported year of birth according to some pre-established density function. Age is then re-derived from the perturbed year of birth.

In the case of the income variables, three different methods are applied to the data. The first one is referred to as the "bottom coding". It addresses the case of the values, positive or negative, that are judged to be close to "0". These identified positive and negative values are respectively used in the calculation of two weighted averages for each province. The identified values are then substituted with their corresponding weighted average. The bottom coded values are finally rounded.

The second method is used for the remaining values. It combines random rounding and the addition of random noise. Such a method makes it difficult for potential intruders to identify the rounding base. The idea is to define a rounding base B and n the number of intervals of length B containing a given value X. One of the bounds of these intervals will be randomly selected to replace X. In order to define these intervals, the following $2n$-1 intervals are considered:

$$I_k = \left[B_{k,0}, B_{k,1}\right], \quad k \in \{1,2,\ldots,2n-1\},$$
$$B_{k,0} = f(X/B)\cdot B - (n-k)\cdot C,$$
$$B_{k,1} = f(X/B+1)\cdot B - (n-k)\cdot C,$$
$$C = B/n.$$

where $f(a)$ represents the integer part of a. Note that n of these intervals include X. Each of these intervals is given the probability

$$\mathrm{P}(I = I_k) = \mathbf{1}_k(X)/n,$$
$$\mathbf{1}_k(X) = \begin{cases} 1 & \text{if } X \in I_k, \\ 0 & \text{otherwise} \end{cases}$$

of being selected. Given a selected interval I_k, X is then randomly substituted with $B_{k,l}$ with the probability

$$\mathrm{P}(S = B_{k,l} \mid I = I_k) = (1 - |B_{k,l} - X|/B)$$

The probability of X being substituted with $B_{k,l}$ is thus given by

$$\mathrm{P}(I = I_k \text{ and } S = B_{k,l}) =$$
$$= \mathrm{P}(I = I_k)\cdot \mathrm{P}(S = B_{k,l} \mid I = I_k)$$
$$= (\mathbf{1}_k(X)/n)\cdot(1 - |B_{k,l} - X|/B).$$

It is easy to prove that $\mathrm{E}(S) = X$. The case for which n = 1 represents the particular case of random rounding.

In order to preserve the analytical value of the data, it is important to make a wise choice of the rounding base B. Too large a value for B would produce data of lesser quality, while too small a value would not sufficiently reduce the risk of re-identification. In the case of SLID, B has been chosen to be proportional to income values.

The third method is referred to as "top coding" and it addresses the disclosure risk entailed by extreme values of income. It consists of the substitution of extreme values with their weighted average. Such a method has not been applied to the data analysed in the next section. Some work still need to be done in order to choose the method of identification of values to be top coded and the aggregation level at which to compute weighted averages (Turmelle, 1999).

6. ASSESSMENT OF THE DISCLOSURE RISK

In this section, we first assess the risk of linkage between two consecutive cross-sectional PUMFs using the income

variables. The 1995 and 1996 data will be used in order to do so. We then evaluate the risk of linkage of the 1995 PUMF to the Income Tax Data File (ITDF). The methodology that is used is based on the work of Grondin, Latouche and Lavigne, 1997 and LaRoche, 1998.

6.1 Linkage of Two Consecutive Cross-sectional PUMFs

In order to evaluate the risk of linkage between the 1995 and 1996 files, different linkage approaches might be used. Here, the chosen methods are: a direct matching approach and a nearest neighbour matching approach. The income variables, the province of residence, the age and the sex are used to carry out these linkage approaches. Disclosure control methods described in section 5 have been applied to data on both files, with the exception of top coding. The risk of linkage is assessed for two provinces, a small one, Prince-Edward-Island, and a large one, Quebec. On the 1995 file, there are 623 individuals from Prince-Edward-Island and 5,411 from Quebec. In the case of the 1996 file, the corresponding numbers are 1,560 and 11,562.

First, the direct matching approach consists of considering as linked a record on the 1995 and a record on the 1996 files for which all their variables are identical on both files. We say that a 1995 record uniquely matches a 1996 record if the 1995 record is unique and the 1996 record is the only one it can be matched to. In addition, only the unique matches between the same individual's 1995 and 1996 records are considered as "serious".

For the second approach to linkage, the nearest neighbour matching approach, some distance calculation between all pairs of 1995 and 1996 unique records is required. We say that a 1995 record i matches a 1996 record i' if the distance d between all their corresponding income source values is less than some predetermined threshold T. More formally, there is a match if

$$d(x_{i,j}, x_{i',j}) \leq T, \quad \forall \text{ variables } j.$$

The distance between two values, a and b, is defined by

$$d(a,b) = |a-b| / \max(|a|, |b|).$$

Unique and serious matches are defined in the same way they were for the direct matching approach. Two different thresholds, $T = 5\%$ and $T = 10\%$, have been used with this approach.

In order to preserve the confidentiality of the PUMFs, it was decided to exclude the results of these record linkages from this paper. However, we briefly discuss the more important results.

The direct matching approach results in a low percentage of unique matches, in the case of PEI as well as in the case of Quebec. Even if higher match rates are obtained using the two scenarios of nearest neighbour linkage, we also consider them as acceptable. Furthermore, we consider that the proportion of unique matches that are not serious is high, which adds a satisfactory level of uncertainty. Given this, we consider that the income variables are not very useful for linking two consecutive cross-sectional files.

6.2 Linkage of the Cross-sectional PUMF and the ITDF

Since a large proportion of the SLID income data are obtained directly from the ITDF, it seems appropriate to assess the risk of re-identification of PUMF records using this file. Consequently, a direct match linkage is conducted in order to assess the risk of linkage between the 1995 PUMF and ITDF. The income variables available on the two files, the province of residence, the age and the sex are used for the linkage. The approach used is similar to the direct match approach described in 6.1. Two linkages are conducted between the two files, one for which the SLID income data have not been perturbed and one for which they have been. The SLID file and the ITDF contain 29,258 and 20,484,469 individuals respectively.

For the same reason as in section 6.1, the results of these linkages are excluded from the paper, but are discussed below.

These linkages show that the application of the disclosure control methods discussed in 5 has a positive effect on disclosure risk. We observe a decrease of 75% in the proportion of unique matches. In addition, a decrease of 65% is observed in the proportion of serious matches among the unique ones. The obtained proportions of unique and serious matches were sufficiently low to consider such linkage not very interesting for re-identification. We also note that the use of such a linkage to add more information to the ITDF would be very inefficient.

To complete this study, a nearest neighbour matching approach is currently being used to link the 1995 PUMF and the ITDF.

7. ASSESSMENT OF THE ANALYTICAL VALUE OF DATA

In addition to its positive effect on disclosure risk, the application of disclosure control methods on data might also have some negative impact on the analytical value of the released microdata. In order to assess the impact the disclosure control methods has on the analytical value of the SLID PUMFs, we compare estimates of different parameters using modified and unmodified data, for different level of aggregation.

Table 1. Relative differences in estimates of totals (%)

Variables	*Can*	*Med. by prov.*	*Max by prov*	*Med. by prov-sex*	*Max. by prov-sex*
Wages & salaries	*0*	*0.1*	*0.22*	*0.1*	*0.44*
Farm self-employ.	*0.3*	*0.82*	*3.38*	*0.93*	*5.07*
Non-farm self-employed	*0.3*	*0.48*	*1.53*	*0.77*	*1.74*
Investment income	*0.2*	*0.5*	*1.11*	*0.48*	*2.3*
Capital gains	*0.6*	*0.77*	*2.81*	*0.79*	*8.74*
Old age pension & GIS, spouse allow.	*0*	*0.11*	*0.43*	*0.17*	*0.74*
Canada or Que pension plan	*0*	*0.12*	*0.3*	*0.21*	*0.57*
Employment insurance benefits	*0*	*0.13*	*0.68*	*0.2*	*1.31*
Social assistance	*0*	*0.35*	*1.02*	*0.51*	*1.59*
Workers' comp.	*0.1*	*0.38*	*0.79*	*1.02*	*5.68*
Pension income	*0*	*0.2*	*0.38*	*0.32*	*1.15*
Other tax. money	*0*	*0.17*	*0.92*	*0.6*	*1.9*
Alimony	*0.3*	*0.35*	*1.46*	*0.83*	*257*
Federal income tax	*0*	*0.18*	*0.49*	*0.33*	*0.69*
RRSP withdrawals	*0*	*0.25*	*0.96*	*0.58*	*2.04*
Prov. income tax	*0.1*	*0.17*	*0.48*	*0.42*	*2.57*
Taxable inv. inco.	*0.1*	*0.22*	*0.99*	*0.62*	*2.13*
Child tax benefit	*0*	*0.34*	*0.97*	*0.76*	*6.55*
Goods/services tax credit	*2.7*	*3.47*	*7.02*	*2.96*	*12.1*
Prov. tax credit	*4.5*	*3.77*	*8.56*	*4.23*	*109*

In table 1, we present the results of a comparison of the weighted totals obtained from modified and unmodified data for different income variables at the Canada level, at the province level and for province-sex domains. The relative differences between the weighted totals obtained from both sets of data are presented at the Canada level. For the last two aggregation levels, the median and the maximum relative differences over the different sub-domains are provided.

At the Canada level, the results show very small differences in the estimates, the exceptions being the variables "goods and services tax credit" and "provincial tax credit". These variables are still identified as showing the largest discrepancies between the modified and unmodified data weighted totals at lower aggregation levels. "Alimony" is also identified as showing the largest discrepancy at the province-sex level. The results in this table show that disclosure control methods applied have a more negative impact on smaller domain estimations, which was predictable.

The large relative differences associated with "goods and services tax credit" and "provincial tax credit" are explained by the fact that these two variables often take values that are closed to "0". This results in a large amount of bottom coded values being assigned for these variables. Since bottom coding results in the same value being assigned for all the records in a province, the frequency of its application might affect the quality of the data. Different solutions such as the modification of the bottom coding threshold for some variables, or the application of bottom coding at lower level could be used to solve this problem. It also seems that random rounding should replace rounding as the final step of bottom coding, in order to preserve the weighted totals at the Canada level. The problem with "alimony" at the province-sex level is also explained by the bottom coding combined with the fact that the number of recipients for some of the province-sex domains is very small.

Comparison of the modified and unmodified data standard deviations and correlations have also been conducted. The results corroborate those obtained in table 1. The highest differences are still observed for "goods and services tax credit" and "provincial tax credit". Apart from these two variables, the results obtained in these comparisons, even if not complete to determine analytical value preservation, are satisfactory.

8. CONCLUSION

The results discussed in the last two sections tend to show that, once the disclosure control methods presented in section 5 have been applied, the income variables are not suitable for the linkage of two consecutive cross-sectional PUMFs. They also show that these disclosure control

methods considerably reduce the risk of re-identification by linking to the ITDF. It was also demonstrated that it is possible to apply disclosure control methods that preserve some of the analytical value of the data, at least at the highest levels of aggregation.

Before the SLID PUMFs are submitted to the Statistics Canada Microdata Release Committee (MRC) for the approval of their release, additional work still needs to be done. In terms of disclosure control methods, the data reduction strategy to be applied to categorical data still needs to be finalized, and the top coding and bottom coding of the income sources still need to be improved. In terms of the risk of linkage of two consecutive cross-sectional PUMFs, the use of the categorical variables is currently being investigated (Franklin, 1999). In terms of re-identification using the ITDF, a nearest neighbour matching approach to linkage is currently used. Some assessment of the risk of re-identification through the use of categorical variables also needs to be done. Such an assessment needs to take under consideration the fact that detailed household and family information is available for any individual on these files. Depending on the different results and on the MRC recommendations, other disclosure control methods, such as data swapping, might also have to be applied to the data. It would also be of interest to obtain more results on the effect the disclosure control methods presented here have on data analysis.

ACKNOWLEDGEMENTS

The authors would like to thank Richard Carter and André Cyr for their helpful comments.

REFERENCES

de Waal, A.G., and Willenborg, L.C.R.J. (1996). A View on Statistical Disclosure Control for Microdata. Survey Methodology, 22, 95-103.

Franklin, S. (1999). Assessing the Risk of Linking Two Cross-Sectional Files. Internal Document in Progress, Statistics Canada.

Grondin, C. (1995). Confidentialité du fichier de microdonnées de l'Enquête sur la dynamique du travail et du revenu (EDTR). Actes du colloque sur les méthodes et applications de la statistique 1995, Bureau de la statistique du Québec.

Grondin, C., Latouche, M. and Lavigne M. (1997). Diffusion de Fichiers de Microdonnées et Confidentialité. Actes du colloque sur les méthodes et applications de la statistique 1997, Bureau de la statistique du Québec.

LaRoche, S. (1998). Rapport sur l'Appariement entre le FMGD du Recensement et le Fichier d'Impôt de Revenu Canada - Fichier de Microdonnées à Grande Diffusion du Recensement de 1996. Confidential Report, Statistics Canada, August 1998.

Latouche, M., Michaud, S., Tremblay, J., and Dibbs, R. (1997). Selection of a Top-up Sample for Cross-sectional Income Estimates. Income and Labour Dynamics Working Paper, Statistics Canada No. 75F0002M 97-04, January 1997.

Lavallée, P., Grondin, C. (1994). Confidentiality of SLID Microdata: A General Approach. SLID research Paper, Statistics Canada No. 75F0002M 94-14, July 1994.

Lavigne, M. and Michaud, S. (1998).General Aspects of the Survey of Labour and Income Dynamics. Income and Labour Dynamics Working Paper, Statistics Canada No. 75F0002M 98-05, March 1998.

Statistics Canada (1997). Survey of Labour and Income Dynamics: Microdata User's Guide, Wave 2. Statistics Canada No. 75M0001GPE, September 1997.

Turmelle, C. (1999). Contrôle de la Confidentialité lors de la Diffusion d'un Fichier de Microdonnées Contenant des Variables Quantitatives avec Valeurs Extrêmes. Internal Document in Progress, Statistics Canada.

Webber, M., Cotton, C., Meere, M., Bishop, K., and Hewer, P. (1999). A Comparison of the Results of the Survey of Labour and Income Dynamics (SLID) and the Survey of Consumer Finance (SCF), 1993-1996. Statistics Canada, No. 75F0002MIE - 990002, April 1999.