Catalogue No. 94-14

CONFIDENTIALITY OF SLID MICRODATA: GENERAL APPROACH

July 1994

Pierre Lavallée, Social Survey Methods Division

Chantal Grondin, Social Survey Methods Division

The SLID Research Paper Series is intended to document detailed studies and important decisions for the Survey of Labour and Income Dynamics. These research papers are available in English and French. To obtain a summary description of available documents or to obtain a copy of any, please contact Philip Giles, Manager, SLID Research Paper Series, by mail at 11-D8 Jean Talon Building, Statistics Canada, Ottawa, Ontario, CANADA K1A 0T6, by INTERNET (GILES@STATCAN.CA), by telephone (613) 951-2891, or by fax (613) 951-3253.

EXECUTIVE SUMMARY

In order to permit longitudinal analyses, the Survey of Labour and Income Dynamics (SLID) is planning to release microdata files. This will bring inevitably a problem related to the confidentiality of the data. Therefore, it is important to develop a general approach by which the disclosure risk of the identity of SLID respondents will be controlled.

The document presents the approach proposed for ensuring that SLID data are confidential. First, a brief overview of some confidentiality principles is given. The classification of the variables in direct identifiers, indirect identifiers and sensitive variables is explained in this section. Second, the processing for confidentiality to be performed on these identifiers and sensitive variables is described. Finally, the general classification proposed for the SLID variables is presented.

TABLE OF CONTENTS

1. INTRODUCTION

The Survey of Labour and Income Dynamics (SLID) covers individuals and households and is used to observe labour market activity and changes in income and family situation over time. The target population of SLID includes all persons living in the provinces of Canada, regardless of age. The Survey excludes the territories, institutions, Indian reserves and military camps for operational reasons. The primary purpose of SLID is to provide longitudinal data, although cross-sectional estimates will also be produced.

In January 1993 a preliminary sample of 15,000 households was selected from two rotating groups from the Labour Force Survey (LFS). This sample forms the basis for the SLID longitudinal sample. In January 1993 the individuals in the sample received a preliminary questionnaire on demographic variables, work and cultural background.

The first interview on labour market activity took place in January 1994. It was designed primarily to identify periods of employment and unemployment of SLID respondents in 1993. In May 1994 interviews were conducted on income and its components for 1993. These labour and income interviews are planned annually for a six-year period.

SLID plans to release microdata files so that longitudinal analyses can be produced. Initially, a microdata file with the results of the preliminary interview and the 1994 labour and income interviews will be released. Subsequently, cumulative microdata files will be released for each of the six waves of the survey (1994 to 1999).

The microdata files contain three types of variables: fixed variables, annual variables and dynamic variables. Fixed variables include characteristics which do not change over time. "Sex" and "age at preliminary interview" can be considered fixed variables. Annual variables include the data from the income questionnaire and all derived variables which provide annual information, such as the indicator "employed or not during the year". Lastly, dynamic variables include all the transition vectors measured by SLID.

Releasing microdata will inevitably lead to the problem of data confidentiality. The purpose of this paper is to propose a general approach to the problem of confidentiality when releasing SLID microdata. We will first give a brief overview of certain principles of confidentiality, where we will explain the classification of variables as direct identifiers, indirect identifiers and sensitive variables. We will then describe how these identifiers and sensitive variables will be processed to protect their confidentiality. Lastly, we will propose a general classification of variables for SLID.

2. PRINCIPLES OF CONFIDENTIALITY

Lavallée (1994) summarized some general principles of confidentiality taken from Schackis (1993). In this section, we will review certain principles taken primarily from Lavallée (1994) which are necessary for the implementation of confidentiality measures for SLID microdata.

There is a disclosure problem if it is possible to produce an unacceptably close estimate of a confidential data item. There are three types of disclosure: disclosure by identification, attribute disclosure and disclosure by inference. Although the last type is fairly important, we will discuss only the first two types here.

Disclosure by identification occurs when a one-to-one relationship is established between released data and a specific respondent. This relationship may be real or only perceived as such, since both situations are harmful to the respondents concerned. It is essential to note that identification is impossible when the respondent is not unique in the population.

Attribute disclosure takes place when various attributes of a respondent are determined with some degree of accuracy from the data that are released. For example, if it is found that all the individuals in a certain age-sex group have a given salary, then the salary of an individual in that age-sex group will automatically be disclosed. The individual will not be identified, but his or her privacy will be threatened simply because certain data were released. Although attribute disclosure is a problem, it is considered less serious than disclosure by identification.

There are three types of confidential data: small counts, cases of predominance and cases of derivation. In the case of small counts, there are too few respondents in some respondent subsets to guarantee their anonymity. In a sample, a small count also corresponds to a small count in the population. In cases of predominance, there can be a large count for a given respondent subset, but one or two respondents are so significant that their percentage contribution to the statistics produced is considered too great. In cases of derivation, there can be a disclosure problem when a number of sources of information, such as several microdata files, are combined. Respondents are then identified by "deriving" their data.

Microdata are organized as records corresponding to each survey respondent. Each record in the microdata file can be divided into three parts:

i) Direct identifiers:

This part includes everything that can clearly identify the respondent, such as name, address, telephone number and individual number in the survey.

ii) Indirect identifiers:

The identifiers in this part cannot be used directly in the sense that a respondent cannot be identified without some knowledge of his or her characteristics. Their value is generally easy to obtain from the respondents. Age, sex, marital status, region of residence and occupation are examples of indirect variables.

iii) Sensitive variables:

This part includes the remainder of the variables in the record. They are referred to as sensitive because they represent private information about the respondent.

Direct and indirect identifiers are also called key variables. The distinction between key and sensitive variables is relatively arbitrary and depends on the desired level of confidentiality for each variable. For example, income may be classified as a sensitive variable if it is felt that income cannot be used in practice to identify an individual. In general, key variables, and particularly indirect identifiers, represent information that is easily available from the public. Sensitive variables, however, are more difficult to obtain because they pertain to the respondents' private lives.

A respondent is said to be unique in the population if his/her indirect identifiers are unique. If a sample is drawn and a unique respondent is selected to be part of the sample, then he/she will also be unique in the sample. The opposite is not necessarily true and, once a respondent has been identified as unique in the sample, a check of the population must be performed. This can be done using administrative data, for example. Obviously, a respondent who is not unique in the sample is not unique in the population.

The risk of disclosure by identification is defined as the probability that an investigator identifies at least one person in the released microdata file. The risk depends on the information provided by the indirect identifiers and the investigator's *a priori* knowledge. The risk of attribute disclosure is the probability that attributes of an individual will be disclosed following the release of indirect identifiers.

3. PROCESSING OF DIRECT IDENTIFIERS, INDIRECT IDENTIFIERS AND SENSITIVE VARIABLES

Confidentiality in SLID will be protected using disclosure control methods. There are two different approaches to disclosure control in the release of microdata. The first is to control disclosure by reducing the information content of the data. This is referred to as disclosure control by data reduction. The second approach consists in changing the data in such a way that identifying a respondent becomes impossible. This approach, called data modification, is generally used after data reduction.

Since SLID is a longitudinal survey designed primarily to measure change, disclosure control methods based on data reduction are preferred over those based on data modification. In addition to perturbing the relationship among the variables of a single wave, data modification methods can in some cases artificially alter the transitions in the data and bias analyses of changes. For longitudinal surveys, these methods must take into account not only the relationship among the variables in a single wave, but the values of previous and subsequent waves. For example, one of the most common data modification methods is random perturbation, where a noise (usually with a mean of 0) is added to the measured values. If generated without taking the other waves into account, this noise can affect the changes from one wave to the next. It can also destroy the consistency in the changes for different variables. Unfortunately, it is generally difficult to modify the data without biasing analyses of changes; this is why efforts to protect the confidentiality of SLID data must be based mainly on data reduction.

Processing for confidentiality of direct and indirect identifiers and sensitive variables should be carried out with a six-year time frame in mind. Confidentiality should not be considered as a progressive operation year after year but rather as a whole. The problem here is that the amount of information on an individual increases over time. If confidentiality methods are applied to the file for the first year, they may no longer be sufficient for the file for all six years. We are faced with a major problem, because the file for the first year has already been released; this is why the six-year approach is important. Note that this approach presents an operational problem because disclosure control methods have to be determined for data which have not yet been collected.

A study of confidentiality over six years is not required for all SLID variables. For fixed variables, processing for six years will be the same as for one year because, by definition, these variables do not change over time. For annual and dynamic variables, the difficulty will depend on whether each variable has been classified as an indirect identifier or a sensitive variable. Annual and dynamic variables classified as sensitive variables will be subjected to disclosure control methods annually, so that they can be processed for confidentiality for a given year rather than for six years. We shall see later that much less effort can be put into confidentiality of sensitive variables than indirect identifiers. Annual and dynamic variables classified as indirect identifiers will have to be studied more extensively using existing data from the Labour Market Activity Survey or tax data. The purpose of the studies will be to identify what degree of change can be expected over a six-year period so that acceptable disclosure control can be determined. Fixed, annual and dynamic variables classified as direct identifiers are discussed in the next section.

3.1 DIRECT IDENTIFIERS

Direct identifiers in the microdata files will be anonymized or else simply removed. The direct identifiers which will be anonymized are those which provide information on the sample design. These include the stratum numbers and component (or replicate) numbers needed to calculate variances. To anonymize these identifiers, their value will simply be replaced by a "meaningless" number.

The other direct identifiers (those which contain no information about the sample design) will simply be removed. These include name, telephone number, address and names of employers.

3.2 INDIRECT IDENTIFIERS

Indirect identifiers are the most problematic from the standpoint of confidentiality. By their very nature, they convey certain information which justifies their presence in the microdata file. Consequently, unlike direct identifiers, they cannot be anonymized or removed without a loss of information. However, the risk of disclosure must be minimized if they are kept in the microdata files.

The following disclosure control methods based on data reduction have been selected for possible application to indirect identifiers: placing restrictions on population size, reduction in detail, top and bottom coding and suppression. Of the disclosure control methods based on data modification, only perturbation by compensation has been chosen.

Placing restrictions on population size consists in setting a minimum population size on which microdata are based. The reason is that as the population becomes larger, the chance of finding two records with the same indirect identifiers increases. We are therefore attempting to define one or more thresholds so that if a population (or subpopulation) is below the threshold, enhanced disclosure control measures are taken. For SLID, we can consider setting a minimum level for the weights used for estimation.

The idea behind reduction in detail is to reduce the amount of information in the released microdata until statistical twins (individuals with the same combination of indirect identifiers) are found in the population. In practice, it is not essential that the twins have all the same indirect identifiers; a subset can be used to characterize them. For example, two individuals can be statistical twins if three of their indirect identifiers are the same. An individual will therefore be unique if the set of all three indirect identifiers is unique. Using this approach, it is possible to reduce the number of indirect identifiers processed at once when processing data for confidentiality (see Lebrasseur, 1994). As mentioned previously, having statistical twins in the sample is a guarantee of having statistical twins in the population; this eliminates the risk of disclosure by identification. There are two versions of reduction in detail: down-scaling and reduction in the number of categories. With down-scaling, numerical values are generally changed to categorical (or qualitative) values which may or may not be ranked.

The purpose of top and bottom coding is to reduce the information content of the microdata by coding the values at the top and bottom ends of the distribution.

These values usually correspond to very small counts within the population, and the risk of identification is therefore high. For the variable "age", for example, a category of persons aged 70 and over can be created. Top and bottom coding is often used following reduction in detail.

Suppression consists simply in removing certain values with a high identification risk from the microdata, as these values pertain to a respondent who probably has no twin in the population. There are two versions of this method: omission of extreme values and removal of the entire record.

Unlike the methods described above, which are based on data reduction, perturbation by compensation is based on data modification. The basic idea behind this method is to provide disclosure control by modifying the microdata in such a way that, once the direct identifiers have been removed (by anonymization), there are no records with a unique combination of indirect identifiers, and perturbations which do not neutralize each other are corrected by corresponding compensations so that the frequency tables are retained. The aim is to create statistical twins artificially by modifying indirect identifiers, while ensuring at the same time that the counts in the frequency tables are the same as before the modifications. For further information, see Schackis (1993).

Once disclosure control methods have been applied to the indirect identifiers, measures will be taken to control the risk of attribute disclosure for certain sensitive variables. The entropy of a sensitive variable is measured for the class of individuals with a given combination of indirect identifiers. If it is low, the sensitive variable has more or less the same value for all the individuals in the class and there is a risk of attribute disclosure. For example, all the individuals in a given age-sex group may have the same income, so that we know the income of the individuals in the group without having to measure it. Such a situation is usually remedied by again applying the method of reduction in detail to the indirect identifiers, in order to increase the entropy of the sensitive variable.

3.3 SENSITIVE VARIABLES

Much less effort is required to protect the confidentiality of sensitive variables than for indirect identifiers. By definition, sensitive variables pertain to individuals' private lives and cannot easily be used for identification purposes. If a variable can be used to identify an individual with or without a reasonable knowledge of the individual's characteristics, it should be classified as a direct or indirect identifier. However, some disclosure control methods will be used to minimize the risk of identification.

The first method will consist in rounding monetary values. The aim of rounding is to avoid a match with files of administrative data such as tax records. For annual data such as total income, values will probably be rounded to the nearest thousand. We propose to round monthly values to the nearest hundred so that they are as consistent as possible with the annual values. Hourly rates for wages, for example, can be rounded to the nearest dollar. Recent studies (see Idehem and Genest, 1994) tend to show that the power of statistical tests does not decrease with rounding of values.

A second method of controlling the risk of disclosure will consist in systematically detecting outliers. It is easy to see that an annual salary of $2 million can make for a fairly high disclosure risk. Each extreme value will be coded or suppressed. For example, let us assume that an annual salary is considered an outlier if it exceeds $200,000. Coding of extreme values will consist in replacing the salary of $2 million with a code indicating that the salary is greater than $200,000. The true value can also be replaced with a statistic such as the mean of the sample values greater than 200,000 or simply with 200,000. This practice is known as Winsorization.

The problem of detecting outliers will lie in how they are defined. For some sensitive variables, univariate methods will be sufficient (frequency tables, box plot, etc.). In other cases, however, each variable will have to be studied in relation to other variables. For example, an income of $100,000 in a small village may be an outlier if the average income in the village is $20,000. In the case of transition variables, often it is not the values themselves but the number of transitions that is abnormal. For example, the marital transition vector may be abnormal if an individual changed status from single to married several times per year.

4. GENERAL CLASSIFICATION OF SLID VARIABLES

4.1 BASIC RULES

4.1.1 Relationship between individuals in the same household

To control the risk of disclosure more effectively, we plan to remove the relationship between individuals in the same household from the output file by reducing the main characteristics of the household to the individual level. This will cause two problems. First, a subset of characteristics that can be used for purposes of analysis will have to be identified. For example, characteristics such as household size and income are generally considered essential variables. It is easy to think of other household characteristics that may only be useful in rare instances, such as the number of children in the household from a previous marriage. Because there can be a huge number of household characteristics, a careful choice will have to be made. Butlin (1994) contains a first draft, but the list could be insufficient if the relationship between individuals in the same household is removed.

The second problem is that we will have to ensure that the household characteristics provided in the output file cannot be used to reconstruct and identify the households. This will be possible if the household characteristic associated with each individual in the output file is unique for a particular household. For example, by associating the "household size" variable with each individual in the output file, it will be possible to reconstruct the only household with a given size. Particular care will have to be taken with "cross-sectional weight", which will probably be unique for each household. This particular characteristic is discussed in the next section. We will also have to pay special attention to family income which, when calculated using the traditional method, (that is, according to the composition of the household defined at a specific point in time), will be the same for all individuals in the same economic family. In many cases, rounding of monetary values and top and bottom coding will solve the problem of reconstructing households.

4.1.2 Household variables

a. Derived household characteristics

With the elimination of the relationship between individuals and households, there is the risk that certain household characteristics added to the file will become indirect identifiers for the individuals in the household. For that reason, all household variables derived from indirect identifiers will themselves be classified as indirect identifiers. For example, a variable which defines the type of family and alludes to the marital status and age of the individual family members will be classified as an indirect identifier.

b. Cross-sectional weighting

The first step in cross-sectional weighting is to calculate the base weight. In a way, this weight represents the sampling weight or the weight before adjustments such as poststratification. The base weight is calculated by taking the mean of the initial weights for the household, which generally correspond to the inverse of the probability of selection of the individuals. This is known as the weight share method and, by definition, it produces identical weights for all members of the same household. The second step in the cross-sectional weighting process is poststratification. The integrated approach is used for this and guarantees a final weight that is unique for all members of the same household.

Because the initial weights tend to differ from one individual to another, after base weighting and integrated poststratification the final weights will quite probably differ from one household to another but be identical for individuals in the same household. It will therefore be possible to reconstruct a household simply by grouping the identical weights together. To avoid this situation, we propose to include more than one household in the weight share calculation. We can show that this method is unbiased provided that the households are grouped independently of the selection process. For example, we can suggest that the households be grouped within the strata-components on the basis of the last digit in the household number; one group would be formed by the households in a stratum-component whose numbers end in 0, another group by the households whose numbers end in 1, etc. Strata-components form a division of the population in relation to the Labour Force Survey.

4.1.3 Geographic variables

To control disclosure more effectively, we are considering removing geographic variables from the output file. This could go as far as removing the province of residence. The basic idea here is that SLID is a Canada-wide survey whose initial aim is not to produce regional or provincial statistics. To take geographic disparities into account, we are considering using qualitative variables derived from the geographic variables. We are thinking of using a "rural/urban area" variable, an indication of the relative poverty of the region, the size of the city of residence, a "family below the poverty line" variable, etc. Note that linguistic variables can be used to roughly identify regions with a high density of Francophones.

An alternative that is being considered in order to retain a geographic variable at the provincial level is data swapping. With this method, data on individuals are swapped within each province so that the regional totals for population counts, for example, remain the same. There are two problems with this method, however. First, the user could not use the "region" variable, for example, for longitudinal studies because the respondents would no longer necessarily have their true region. Data swapping is actually interesting only from a cross-sectional point of view, which is a secondary priority for SLID. Moreover, it would be easy for the user to use the "region" variable carelessly and draw erroneous conclusions. The second problem is that data swapping is difficult to do in practice. The problem does not lie in swapping the regions themselves, but in ensuring that the population counts are maintained. Because of these two problems, we feel that data swapping for the "region" variable is not a very attractive alternative.

4.1.4 Transition vectors

SLID measures a number of dynamic variables in relation to work, educational activity, unemployment and periods of social assistance, for example. These variables appear in the output file as transition vectors. Transition vectors will not be considered confidential data, because it can be very difficult or even impossible to relate them in practice to a specific individual. Consequently, unless some transition vectors are completely extreme or strange, they will not be subject to disclosure control, and will be classified as sensitive variables rather than indirect identifiers. This decision alleviates the difficulty of disclosure control considerably by substantially reducing the number of indirect identifiers.

4.1.5 Fields containing text

Most fields containing text will be removed from the output file, particularly if the information in the text increases the risk of disclosure. These variables include the interviewer's notes; name of employer; kind of business, industry or service; the description of the kind of work done; main activities or duties at work and major field of study. Most of this information will be coded and only the coded values will be kept in the output file. Needless to say, the coded values will be subject to the SLID confidentiality process.

4.2 GENERAL CLASSIFICATION OF VARIABLES

4.2.1 Direct identifiers

In general, the following will be classified as direct identifiers:

- individual and household identifiers from the LFS and SLID (to be removed).
- the respondent's surname, given name and work and home telephone numbers (to be removed).
- the contact person's surname, given name, address, telephone number and identifier (to be removed).
- the name of the respondent's employer (to be removed).
- information from the sample frame: primary sampling unit (PSU), rotating group, stratum, component, assignment number, interviewer number (to be anonymized).

4.2.2 Indirect identifiers

In general, the following will be classified as indirect identifiers:

- geographic variables (to be removed, except perhaps province of residence).
- the respondent's date of birth, age, sex, preferred language, language of interview and marital status, and all other related variables.
- the respondent's relationship to the other household members.
- the employer's name, the type of work and the industry and occupation codes for all jobs.
- variables pertaining to type of dwelling (from the LFS).

- variables pertaining to ethnic origin, country of origin and mother tongue.
- number of children.
- variables pertaining to education: major field of study, attainment of a degree in medicine or optometry, a doctorate, etc...
- household variables derived from indirect identifiers.
- interviewer check items pertaining to indirect identifiers: for example, check if the respondent is over 70 years of age (to be removed).

4.2.3 Sensitive variables

All other variables -- that is, those pertaining to people's private lives -- will be classified as sensitive variables. These include variables not mentioned above related to labour and income and certain household variables not related to indirect identifiers (such as family income).

5. BIBLIOGRAPHY

Butlin, G. (1994). *Household and Family Variables*. SLID Research Paper Series 94-06, 1994.

Idehem, I., and Genest, C. (1994). *L'utilisation de données regroupées et son impact sur l'inférence en analyse de la variance*. Proceedings of the 1994 symposium on statistical methods and applications (to be released), Quebec Bureau of Statistics. Quebec City, 1994.

Lavallée, P. (1994). *Confidentialité des microdonnées de l'EDTR: Approche générale*. Internal document. March 11, 1994.

Lebrasseur, D. (1994). *Différents aspects de la méthodologie dans la production des fichiers de microdonnées à grande diffusion dans le cadre du recensement de 1991*. Proceedings of the 1994 symposium on statistical methods and applications (to be released), Quebec Bureau of Statistics. Quebec City, 1994.

Schackis, D. (1993). *Manual On Disclosure Control Methods*. Eurostat - D3. Luxembourg, September 1993.