# RESEARCH INTO THE POSSIBILITY OF RELEASING HIERARCHICAL PUBLIC USE MICRODATA FILES FOR THE CANADIAN CENSUS OF POPULATION

Jean-René Boudreau
Rossana Manríquez

Statistics Canada
R.H. Coats Building, 15th floor ( Tunney's Pasture ) Ottawa, Ontario, K1A 0T6, Canada
jean-rene.boudreau@statcan.ca / rossana.manriquez@statcan.ca

**Abstract**

A microdata file is the set of answers obtained through a survey. For household surveys or population censuses, files of this type are hierarchical if the data on all the persons in the households sampled or enumerated are present. The statistical agencies ensure that this data remain anonymous by eliminating all possibility of identification. A number of countries produce hierarchical files of microdata from their censuses. Canada, however, does not. As a matter of fact, Canada has disseminated microdata files since 1971 but these files have always been designed without incorporating the complete hierarchy of households. We are assessing here the protection of statistical confidentiality for the 2001 Census, that is, we are assessing whether it is easy or very difficult to identify individuals from data alone. We use two measurements of statistical confidentiality protection: the conditional probability of uniqueness and the conditional probability of exact matches. We apply these two measurements to a set of records from the 1996 census for various groups of variables. On the basis of the results obtained, we conclude that publication of a public use hierarchical file significantly reduces the protection of statistical confidentiality.

**Keywords**: Risk of disclosure; Public use microdata; Confidentiality; Anonymised records (SAR);

## §1. Introduction

The Census Public Use Microdata Files (PUMFs) contain samples of anonymous responses to the Census questionnaire. Those files are unique among census products in that they give users access to non-aggregated data. We have been disseminating those files for each census since 1971. Three files are available: the Individuals File, the Families File, and the Households and Housing File. Each of the files contains information on most of the census subject matter on approximately 3% of the Canadian population. There is information about demography, schooling, income, labour force activity, housing, family, language spoken and ethnic origin. This data is available for the provinces, territories, selected census metropolitan areas and selected census subdivisions. In order to protect the confidentiality of the information provided, special measures were taken. For example, these files do not contain any information on the same people. This ensures that the content of the three files is controlled.

A hierarchical file is a household file for which all the data for all the persons in the households are present. Up to now, the Canadian PUMFs are not hierarchical, but have a partial hierarchy. For instance, in the individuals file, there is information on the family structure and on the household income. We began a research program to re-assess the possibility of distributing hierarchical microdata for the 2001 Census. This re-assessment is necessary if we keep in mind the following facts:

(a) There are three post-censal surveys that could interfere with the PUMFs sample: the survey on aboriginal peoples, the activity limitation survey and the survey on ethnic diversity. These surveys take their samples from census data. The individuals included in their target populations often have characteristics that are uncommon. The census microdata file designers must ensure that none of these surveys' samples overlap with theirs: otherwise, the total content published would no longer be controlled.

(b) Public use census microdata files are a set of three files of microdata drawn from three universes: individuals, families and households. The designers of these files also ensure that the three files will not overlap in the sense that the family and household of a chosen individual in the individuals file will not be drawn from one of the other microdata files and vice versa. The way to guarantee that these three samples do not overlap since the 1991 Census, although efficient, does not optimize the use of the records. If this method of drawing samples is used in 2001, too many records could be

eliminated at the time of the first two first draws, leaving an insufficient number of records for the last draw. Creating a hierarchical microdata file would solve this problem.

(c) The Canadian scientific community wants to be able to use the household-family-individual hierarchy in their analyses. Clearly, the presence of hierarchy in public use microdata would make this possible.

(d) Finally, the international scientific community wants to create microdata banks for various countries in order to perform multinational analyses. For the vast majority of countries where microdata files exist, these files are hierarchical.

For all these reasons, it became more urgent to base the decision regarding hierarchical public use microdata files on acceptable scientific arguments. The text that follows is the outcome of this research.

**§2. Underlying theory**

The decision regarding public use of microdata files or any other product must be based on disclosure risk. By definition, disclosure risk is the plausibility, degree of confidence, probability or even the impression that an intruder[1] can, from a product, determine the answers given by an identifiable individual. In terms of a microdata file, risk must be defined in the following terms:

(a) The probability or frequency of unique records in the population found among the unique records in the sample. Uniqueness is defined by the intersecting of key variables (variables that are often found in other microdata files). This statistic will be called the "conditional probability of uniqueness".

(b) The probability of obtaining exact matches, given that one is able to twin records on a one-to-one correspondance. Match variables are the key variables. Basically, an intruder asks himself if the matches obtained are from the same person only after having matched a file containing names, addresses, etc. with the unique records in the file of public use microdata. This statistic will be called the "conditional probability of exact matches".

Traditionally, we have always tried to assess the conditional probability of uniqueness as a means to measure disclosure risk. This probability is, however, difficult to determine because the number of unique elements in the population must be assessed (which is not possible from the sample alone). On the other hand, the conditional probability of exact matches is closer to what we should be aiming at: an intruder is only interested in the likelihood of exact matches among the matches he/she obtained. We are going to use these two measures to compare disclosure risk of public use microdata files with a hierarchised microdata file, but we are going to treat as more important, or place more emphasis on, the conditional probability of exact matches.

In this text, the term "content" has a very precise meaning. We define the content of a microdata file as the description of the population in terms of the cross tabulation of the key variables. This table includes $m$ cells of size $N_1$ , $N_2$ , … , $N_m$. To describe the population in terms of a table amounts to clarifying the number of cells with only one element ($U_1$ the number of unique elements), the number of cells with only two elements ($U_2$ the number of twins), with three elements ($U_3$ the number of triplets), etc. Generally speaking, $U_j = \text{card}( \{ k : N_k = j \} )$. The vector ($U_1$ , $U_2$ , … , $U_N$) gives the content of the population. The measure of disclosure risk, either by (a) or by (b), depends on the content of the file under observation, its sampling ratio and a sample. We now give the formulas associated with the mean over all possible samples of the definitions described in (a) and (b) for a population of size $N$, a sampling size $n$, a sampling ratio $f$ and under the assumption of simple random sample.

**Conditional probability of uniqueness**

$$f U_1 \left( \sum_{i=1}^{N-n+1} i U_i \binom{N-i}{n-1} \Big/ \binom{N}{n} \right)^{-1}$$

**Conditional probability of exact matches**

$$\sum_{i=1}^{N-n+1} i U_i \binom{N-i}{n-1} \Big/ \sum_{i=1}^{N-n+1} i^2 U_i \binom{N-i}{n-1}$$

[1] An intruder is someone who seeks to disclose information of an individual from a published statistic.

**§3. Description of Data**

The data comes from the 1996 Census. We have selected records from each of the five regions of Canada in private households. (If an individual is in our data, all the members of the household are there as well.) The number of records in each region is approximately 250,000. In order for a region to be identified in the microdata files, it must have at least 250,000 inhabitants except for the small provinces. Table 1 summarizes the database used in our research.

**Table 1 : Number of Households in the Database According to Region and Size of Household**

| Region | Total | Size 1 | Size 2 | Size 3 | Size 4 | Size 5 | Size 6 and over |
|---|---|---|---|---|---|---|---|
| **Total** | 474,275 | 112,801 | 151,003 | 81,020 | 80,649 | 32,673 | 16,129 |
| Atlantic | 89,818 | 16,527 | 27,843 | 17,934 | 17,745 | 6,858 | 2,911 |
| Quebec | 99,691 | 25,794 | 31,540 | 18,033 | 16,827 | 5,780 | 1,717 |
| Ontario | 97,179 | 23,277 | 32,198 | 16,487 | 16,598 | 6,178 | 2,441 |
| Prairies | 92,336 | 23,403 | 27,820 | 14,454 | 14,606 | 7,110 | 4,943 |
| Pacific | 95,251 | 23,800 | 31,602 | 14,112 | 14,873 | 6,747 | 4,117 |

We will make several comparisons of content based on individual subject matter. These subjects[2] are:

(a) demography;
(b) ethnic origin and immigration;
(c) labor force activity;
(d) schooling;
(e) sources of income;
(f) language spoken.

In our calculations, we used a sampling rate of 3%, which corresponds roughly to the rate used in the microdata files. For the hierarchical file we used the database of households ( 474,275 ) and for the non-hierachical file, we used the database of persons ( 1,250,488 ). We considered those databases as being the population.

**§4. Hierarchised content associated with a content**

We want to compare the disclosure risk of a file with the level of risk when "hierarchising" its content. A hierarchised microdata file is basically a file of households containing all the information on the persons living in it. We have to find a way to put all the information available on the individuals and families at the household level.

In order to define the hierarchised content associated with a content, we begin by "hierarchising" a variable. A variable in the universe of individuals defines a variable at the household level by using as code the list of the original codes of the variable for all members of the household in a particular order. Thus, if the genders of the individuals in a household are, in descending order of age, {MALE, FEMALE, FEMALE, FEMALE}, then this household's hierarchised sex variable has the code MFFF, with M designating MALE and F, FEMALE. The definition of a hierarchised variable depends on the order of the individuals chosen. In order to make comparisons, we always take persons in the following order: first, the primary household maintainer, then the person married to or living in common law with the primary household maintainer (if there is one) and then the other persons in the household according to the subject being studied. Thus, for demographic variables, the persons other than the primary household maintainer and his/her partner are placed in order according to age, sex,… while for variables related to ethnic origin, these persons will be placed in order according to their ethnic origin and citizenship. After having defined the hierarchised variables, the associated hierarchised content is the content obtained by taking these hierarchised variables to which the structural variables of the household are added. By structural variables, we mean household size, family relations, number of mainteners and so on. Those variables add a partial hierarchy to the file.

[2] Mobility and housing are not dealt with in this document but are present in the individuals file.

**§5. Risk analysis for demographic variables**

In the 1996 individuals file, we find the following demographic variables: age; sex; legal marital status; marital status indicator-historical comparability. We manage to have exactly the same content as the individuals file for demographic variables.

Table 2 gives the conditional probabilities of exact matches and uniqueness. The column entitled "Non-hierarchised" shows those probabilities for which only structural variables present in the file are added to the demographic variables to define uniqueness. Information on other persons in the household is not taken into consideration, even if available. The column entitled "Hierarchised" shows probabilities of demographic content hierarchised with all structural variables. Calculations for the column entitled "Non-hierarchised" are made at the individual level and calculations for the column entitled "Hierarchised" are made at the household level.

The data should be interpreted as follows. Let's assume that your neighbour's household (in Ottawa) is made up of four persons. You may know these persons well enough to know their demographic and structural variables. If you find this information on a hierarchical census file and there is only one household with the same data, you may reckon more than nine times out of ten (93.93%) that these persons identified in the file are in fact your neighbours.

Since the chance of knowing demographic and structural characteristics is not an unusual occurrence (the majority of Canadians know the exact characteristics of some households of at least 4 people), that makes the identification of records very plausible since nearly all these households are effectively unique in the file.

**Table 2 : Conditional Probabilities of Uniqueness and Exact Matches for Demographic Variables.**

| Region | Household size | Structure | | | |
|---|---|---|---|---|---|
| | | Non-hierarchised | | Hierarchised | |
| | | Unicity | Match | Unicity | Match |
| **Total** | | | | | |
| | **Size 1** | 0.40 % | 2.44 % | 0.40 % | 2.44 % |
| | **Size 2** | 6.49 % | 4.88 % | 30.15 % | 9.84 % |
| | **Size 3** | 15.70 % | 6.68 % | 64.82 % | 53.58 % |
| | **Size 4** | 21.12 % | 7.26 % | 82.13 % | 79.91 % |
| | **Size 5** | 26.71 % | 8.76 % | 98.88 % | 98.88 % |
| | **Size 6 and more** | 38.54 % | 13.28 % | 99.83 % | 99.81 % |
| **Ontario** | | | | | |
| | **Size 1** | 1.48 % | 3.16 % | 1.48 % | 3.16 % |
| | **Size 2** | 10.99 % | 6.32 % | 37.43 % | 16.06 % |
| | **Size 3** | 21.30 % | 7.06 % | 84.24 % | 81.59 % |
| | **Size 4** | 24.52 % | 6.80 % | 94.05 % | 93.93 % |
| | **Size 5** | 30.67 % | 7.53 % | 99.84 % | 99.84 % |
| | **Size 6 and more** | 44.12 % | 14.81 % | 99.92 % | 99.92 % |

Can these probabilities be reduced? There are four ways:

(a) Group the geography. The fourth column of Table 2 gives the probabilities of exact matches for one geography and for the whole database. When the geographies of the regions are no longer used, the probabilities decrease slightly. It is necessary to realize that there are over 1,000,000 persons included in the database used. This has had no noticeable effect. This is not the way to reduce the amount of disclosure risk.

(b) Group the variables. For example, for households of 5, we have obtained the results shown in Table 3. It can be seen that the degree of risk is reduced significantly. On the other hand, it is necessary to group all the regions as well as group age into 10 year ranges in order to find a degree of risk similar to that of the 1996 individuals file. The file with these groupings will have lost nearly all usefulness, as age is an essential variable. Demographic variables really cannot be removed. Age, sex, marital and common-law status are essential. By the same token, structural variables cannot be removed either, as they are implicit in a hierarchical file.

(c) Reduce the sampling ratio. This method reduces the conditional probabilities of exact matches or uniqueness only if the content of the population is not degenerated (a content is degenerated when the $U$'s are all 0 except for very small $i$ indices). Now the content studied here is degenerated. For example, we have in the file, data on 32,673 households of 5. The content for these households is $U_1$ = 32,297, $U_2$ = 185 and $U_3$ = 2. With this type of content, the reduction of the sampling ratio has no effect on the degree of risk.

(d) Introduction of noise in a sub-set of records. The methods known to the authors are: data swapping, suppression of values, fluctuation of variables (that is, adding white noise to continuous variables such as sources of income). If we modify the values of variables too much, it will certainly affect the quality of the analyses. It is our opinion that these methods must introduce so much noise in order to reduce the risk, that the file's usefulness will be decreased, if not eliminated entirely.

**Table 3 : Conditional Probabilities of Exact Matches for Persons in Households of 5 for Different Age Groupings.**

| Region | Groupings | | | |
|---|---|---|---|---|
| | None | By 2 | By 5 | By 10 |
| **Total** | 98.88 % | 80.14 % | 20.13 % | 11.17 % |
| **Atlantic** | 99.80 % | 96.34 % | 49.13 % | 19.17 % |
| **Quebec** | 99.73 % | 94.20 % | 42.38 % | 17.06 % |
| **Ontario** | 99.84 % | 94.31 % | 48.26 % | 18.43 % |
| **Prairies** | 99.59 % | 93.37 % | 47.50 % | 18.42 % |
| **Pacific** | 99.63 % | 95.52 % | 52.70 % | 21.53 % |

Thus, for households of 4 and more (which represent about 27% of households and 47% of the individuals in our database), the possibility of identification is nearly certain and the usual methods for reducing degrees of risk do not work.

**§6. Risk analysis for other subject matter areas**

We have done a similar analysis with other subject matters included in the 1996 individuals file. These subjects are ethnic origin and immigration, labour force activity, schooling, sources of income and language spoken. We have tried to reproduce the same content. If this was not possible due to the difficulty of deriving the latter, the content that we used for this research was always less refined than that published. This implies that the probabilities we give are smaller than the actual probabilities.

**Table 4 : Conditional Probabilities of Uniqueness and Exact Matches for some subject matter, Households of 3.**

| Subject matter | Structure | | | |
|---|---|---|---|---|
| | Non-hierarchised | | Hierarchised | |
| | Unicity | Match | Unicity | Match |
| **Ethnic Origin and Immigration** | 12.49 % | 5.91 % | 44.94 % | 14.40 % |
| **Labor Force Activity** | 16.54 % | 7.65 % | 69.23 % | 32.42 % |
| **Schooling** | 12.60 % | 6.43 % | 55.09 % | 21.67 % |
| **Income Sources** | 47.56 % | 14.86 % | 75.91 % | 40.76 % |
| **Language Spoken** | 15.69 % | 6.46 % | 35.61 % | 9.59 % |

The probabilities labeled "Non-hierarchised" are calculated based on the definition of uniqueness which only takes into account structural variables and variables of the various subject matters used. These probabilities measure the disclosure risk of the individuals file. As for the conditional probabilities of exact matches, they are all less than, or close to, 10% except for those concerning sources of income. We note that these latter are only given for the purpose of information because they depend on the arbitrary defining of income ranges. By preserving conditional probabilities of exact matches below 10% (or very close to it), we feel that the disclosure risk of the individuals file is fairly well controlled.

This is not the case for a hierarchical file with the same content. It is not really ethnic origin or more generally, sociocultural variables that cause the most problems. In fact, people living together have very similar sociocultural traits, thus decreasing the probability of exact matches. Demographic variables are the ones that are so problematic as well as those related to the

labour force activity, schooling and sources of income. The high level of their probabilities of exact matches makes even their publication dangerous (assuming that there is always an intruder in the population). We have shown in the section on demographic variables that in order to reduce these probabilities to an acceptable level (e.g.: similar to those of the non-hierarchical structure), it would be necessary to reduce the file's usefulness so much that no one would be interested in getting it.

An intruder will certainly use demographic variables in conjunction with those of another subject matter to define uniqueness. We give in Table 5, the probabilities of exact matches for the subjects studied but to which we have added, to define uniqueness, the demographic variables. In this case all the subjects become problematic.

**Table 5 : Conditional Probabilities of Uniqueness and Exact Matches for some subject matter augmented with demography, Households of 3.**

| Subject matter | Structure | | | |
|---|---|---|---|---|
| | Non-hierarchised | | Hierarchised | |
| | Unicity | Match | Unicity | Match |
| **Ethnic Origin and Immigration** | 33.15 % | 12.37 % | 95.73 % | 95.02 % |
| **Labor Force Activity** | 45.57 % | 19.85 % | 99.22 % | 99.11 % |
| **Schooling** | 39.02 % | 15.32 % | 98.97 % | 98.88 % |
| **Income Sources** | 64.49 % | 26.50 % | 99.81 % | 99.80 % |
| **Language Spoken** | 30.82 % | 9.74 % | 86.40 % | 80.68 % |

## §7. Conclusion

In this text we have tried to determine a measurement of the disclosure risk of hierarchised microdata with variables similar to those in the individuals file of the 1996 Census. We have used two means of measurement. The first, the conditional probability of exact matches, was developed quite recently by Elliot in Great Britain (Elliot[2]). It estimates the probability that if there is a match, it is exact. The second, the conditional probability of uniqueness, described in Boudreau [1], estimates the probability that a unique element in a sample is unique in the population. These two quantities measure the possibility of identifying the records of an individual.

We have shown that if we publish a hierarchical file with only demographic and structural variables, the possibility of identifying the individuals, for households of four or more, is almost certain. The variables related to the labor force activity, schooling and sources of income are equally unsafe. The methods for reducing the degree of disclosure risk are ineffective for these households. Making the disclosure risk similar to that of the individuals file would reduce the file's usefulness too much.

## §8. References

[ 1 ] : Boudreau, Jean-René. 1996. "Assessment and Reduction of Disclosure Risk in Microdata Files Containing Discrete Data". *Symposium '95, From Data to Information: Methods and Systems: Proceedings.* Ottawa. Statistics Canada. 143-153.

[ 2 ] : Elliot, Mark. 2001. "Disclosure Risk Assessment", In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Pat Doyle, Julia I. Lance, Jules J.M. Theeuwes and Laura V. Zayats (eds). Amsterdam. North-Holland Elsevier Science. 75-90.