Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

The 2011 National Household Survey public use microdata file methodology: How to balance the requirement for more information and the requirement for low risk of disclosure in the microdata

Chunxiao (William) Liu and François Verret[1]

Abstract

The 2011 National Household Survey (NHS) is a voluntary survey that replaced the traditional mandatory long-form questionnaire of the Canadian census of population. The NHS sampled about 30% of Canadian households and achieved a design-weighted response rate of 77%. In comparison, the last census long form was sent to 20% of households and achieved a response rate of 94%. Based on the long-form data, Statistics Canada traditionally produces two public use microdata files (PUMFs): the individual PUMF and the hierarchical PUMF. Both give information on individuals, but the hierarchical PUMF provides extra information on the household and family relationships between the individuals. To produce two PUMFs, based on the NHS data, that cover the whole country evenly and that do not overlap, we applied a special sub-sampling strategy. Difficulties in the confidentiality analyses have increased because of the numerous new variables, the more detailed geographic information and the voluntary nature of the NHS. This paper describes the 2011 PUMF methodology and how it balances the requirements for more information and for low risk of disclosure.

Keywords: Large-scale survey, individual and hierarchical public use microdata files, multiplicity, uniqueness prediction, random rounding and residual disclosure.

1. Introduction

The Canadian census of population targets the whole Canadian population and is conducted every five years. Historically, the census consisted of a mandatory long-form questionnaire sent to one in five households and a mandatory short-form questionnaire sent to the rest of the population. In 2011, the census consisted of only the mandatory short form, while the long-form data were collected with a voluntary survey, the National Household Survey (NHS). The survey sampling fraction was 30%, up from 20%, but the NHS design-weighted response rate was 77%, while for the 2006 long form it was 94%.

The NHS public use microdata files (NHS PUMFs) have the same target population as the NHS; they are created from the NHS respondent data of 6.7 million individuals. Two NHS PUMFs are released: the hierarchical PUMF (HPUMF) and the individual PUMF (IPUMF). Both are composed of records of individuals. The main difference is that the HPUMF contains hierarchical information linking individuals within families and households. The sampling fractions of the HPUMF and IPUMF are 1% and 2.7% of the target population, respectively.

A PUMF contains information on individual records and many variables. Its abundant information gives it analytical power, but it comes with potential disclosure risks. As required by Canada's *Statistics Act* and by the Fundamental Principles of Official Statistics put forward by the United Nations Statistical Commission, data collected by the Statistics Canada are to be kept strictly confidential. Since there is a tension between the requirements of providing more data and maintaining low disclosure risk, we must find a balance during the creation of a PUMF to satisfy the two requirements.

[1] Chunxiao (William) Liu, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Canada, K1A 0T6, Chunxiao.Liu@statcan.gc.ca.
François Verret, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Canada, K1A 0T6, François.Verret@statcan.gc.ca.

In this paper, we describe the methodology for the creation of the IPUMF and the HPUMF of the 2011 NHS. In section 2, we give basic considerations in developing the PUMFs. In section 3, we present sampling and weighting methodologies. In section 4, we explain the confidentiality analysis method, and in section 5 we describe data treatments. Sections 3, 4 and 5 focus on the IPUMF. However, we applied the same methodology in the development of the HPUMF, with differences outlined in section 6. In the conclusion, we recapitulate how we balanced the requirements in the creation of the PUMFs. Information on the content of the 2006 and 2011 PUMFs can be found in the respective User Guides (Statistics Canada, 2010, 2011, 2014a and 2014b).

2. Basic considerations

Firstly, privacy protection is the top priority when producing a PUMF. Respondents should not be identifiable either directly or indirectly. This paper focuses on how to deal with potentially indirect identifiers (for example, geography, sex, age and family structure). Crossing, or combining, these identifiers could effectively identify an individual in the population. Secondly, the content of the PUMF should be comparable with that of the previous cycle's PUMF. For 2011, we had to consider some new variables and more detailed categories of variables that were added that year. Thirdly, the starting point for this cycle was the methodology used in the 2006 production cycle, since it has been accepted by Statistics Canada's Microdata Release Committee and has been well received by data users. However, we had to make improvements to account for the non-response rate, which is higher than that for the census long form, and varies a lot across sub-populations, and to account for the new content. Finally, since we created more than one PUMF from the same NHS data, an effort should be made to ensure that no individual is in more than one microdata file. Because data treatment is done independently for each PUMF, users could link two files to reveal more information on individuals included in both than we intended to provide, potentially creating a disclosure risk. On the other hand, postcensal surveys use the NHS data for their sampling frames, so we cannot avoid overlap with the NHS PUMFs without introducing significant bias. Consequently, we should treat overlapping individuals more forcefully.

3. Sample selection and weighting

Privacy protection begins with the sampling and weighting procedures. For protection, we use only a fraction of the NHS data of 6.7 million individuals to create each of the PUMFs. Three PUMF samples were drawn: the IPUMF sample, the HPUMF sample, and a sample reserved for the possible production of a third PUMF intended to do international comparisons. The methodology of this third PUMF is not described in this paper. The sampling design for drawing each of the PUMF samples has two phases. In the first phase of sampling, we divided the frame of NHS respondent households into three sub-frames. Each sub-frame was used for drawing a PUMF sample in the second phase of sampling. The sizes of the sub-frames are proportional to the PUMF sampling fractions. The three PUMF samples are therefore non-overlapping, and the second-phase sampling fractions of the sub-frames are the same for all three PUMF samples.

In the first phase of sampling, each sub-frame should be well-balanced in terms of geography and household size. For this purpose, we sorted households geographically and by household size and applied a systematic sampling method. This process is called implicit stratification. Since the sub-frames are samples of NHS data, we adjusted the NHS weights according to the sampling design so the sub-frame weights would correctly represent the population. In the second phase of sampling, as in the first phase, we drew well-balanced PUMF samples from the sub-frames using systematic sampling and appropriate sorting. Before drawing the IPUMF sample, we sorted data by province (or territory), urban–rural indicator, gender, age group and ethnic-origin group. On top of that, to improve statistical efficiency and to have a self-weighted PUMF sample, we used a probability-proportional-to-size (PPS) systematic sampling method, where the size measure was the adjusted NHS weights.

A PUMF sample with identical weights is desirable because it does not make any individual stand out in terms of the PUMF weights. However, it was not possible to select a self-weighted sample because the size measure could have been bigger in some cases than the sampling interval of the PPS systematic sampling schema. The closest we could come to a self-weighted sample was to select with certainty the individuals whose size measure is too big, and then select a self-weighted sample from the remaining individuals using the PPS systematic sampling method (Särndal, Swensson, and Wretman 1992).

Since all household members have the same size measure, if a member was selected with certainty in the second-phase sample, all other members of the household were also included in the sample. In the IPUMF sampling, for protection, some of the individuals selected with certainty in the second-phase sample were sub-sampled and excluded from the final file. To make sure the final IPUMF sample contained 2.7% of the target population, we raised the actual sampling fraction for the IPUMF in the first and second phases appropriately.

For variance estimation, we used the method of dependent random groups (Wolter 1985). We created eight groups and made a replicate weight available to the users for each group. Although this method tends to be conservative, it has the advantage of being very simple to implement, since only the first phase of the multi-phase design (i.e., the first phase of the NHS design) has to be replicated. It is also very simple for the user to perform variance estimation using replicate weights.

4. Confidentiality analyses

The goal of the confidentiality analyses is to identify any individuals with potential risks of disclosure in the microdata file. Could someone identify his or her neighbour by crossing the information provided in the file? If that possibility exists, we should reduce the neighbour's data in data treatment. At first glance, it would appear evident to use the PUMF sample for the analyses instead of the entire set of NHS data, since it is smaller, which makes analyses simpler, and since only the sample will be published. However, an individual who stands out in the sample does not necessarily stand out in the population. To reduce the uncertainty, we used the full NHS data. The core of the analyses is three-way table analyses, which cross information for three variables.

4.1 Analysis of the NHS data

From over 120 variables in the IPUMF, we identified about one-third as indirect identifiers, and included them in the analysis. Indirect identifiers represent characteristics of an individual that could be used to identify him or her, such as age, sex, place of birth, ethnic origin, education and income. The list of indirect identifiers comes from an accumulation of subject-matter experience over the years. To update the list, we focused more on the new variables. It is safer to put them in the analysis than to leave them out.

Among the identifiers, some are closely related. In these cases, it is more efficient to combine them as one variable in the analysis. For example, place of birth (POB) and place of birth of parents (POBF and POBM) can be concatenated to form the variable POB_NEW. We applied the combination method to variables on demography, education, labour, and housing and shelter costs. Effectively, this process reduced to 22 the total number of variables to be analyzed. We primarily analyzed categorical variables, but we also analyzed some numerical variables, such as total household income and housing and shelter costs. We categorized the continuous values before the analysis.

Geographic location and type of household or family are essential in identifying an individual. This information needs to be treated carefully as it tends to be known among people from the same neighbourhood. In the PUMFs, the geography information can be detailed to large cities. The family and household can correspond to, for example, a lone parent with children, or a couple with or without children. For the IPUMF, we used this information to define 315 domains (created by crossing the 35 identifiable geographies with 9 universes defined by demographic, census family and household characteristics). We conducted the three-way table analyses independently for each domain and generated over 400,000 three-way tables.

The purpose of the three-way tables is to find any individual record alone in a table cell. A record alone in a table cell is called a *unique case*. Since an individual record can be involved in more than one unique case, the number of unique cases for a record is called the *record multiplicity*. A variable can also be involved in more than one unique case for a given record. The number of unique cases of a variable for a given record is called the *variable multiplicity* for the record. A variable is said to be the worst one for a given record if its variable multiplicity is the highest among the variables in the analysis. For example, consider an analysis with five sensitive variables: A, B, C, D, and E. If record #1 is a unique case in three-way tables ABC, ABD and ACE, its record multiplicity is 3. Its variable multiplicity is 3 for variable A, 2 for B and C, and 1 for D and E. Variable A is the worst one for record #1.

Since the NHS collects data from only part of the population, a unique case in the NHS might not be unique if we could consider all individuals whose data were not collected by the NHS. Uniqueness at the population level should thus be predicted based on the sample information before data treatment is applied. Records predicted as being unique in the population are considered at risk of disclosure and should have their data reduced.

4.2 Uniqueness prediction in the population

Stating that "a unique case in the sample remains unique in the population" is equivalent to saying "all individuals non-collected by the NHS fall outside the table cell where a unique case is observed in the sample data." We need to predict this event with only the sample data. The table cell of a unique case identified in the sample is called a unique cell. Because there is only one observation in the unique cell, prediction of the event in the design-based setting is either impossible or very imprecise (since the unique cell is never a design stratum). Thus, we assume a superpopulation model for the non-collected individuals, i.e., the number of non-collected individuals falling into the unique cell is a random variable following a binomial distribution with probability p .

The probability that a unique case in the sample is also unique in the population can be estimated by $(1-\hat{p})^{N_{non-collected}}$, where $\hat{p}$ is an estimator of p based on the sample. Estimation of p is not straightforward, since by the nature of the analysis only one individual can contribute to the estimation of the numerator of the proportion. Under the assumption that the non-collected individuals follow the same distribution as the finite population, one could be tempted to use the weighted proportion as an estimator. However, although the estimator could be unbiased, it is not reliable because of its large variance. Furthermore, instability of the estimator makes the inference described below unreliable. On the other hand, if one seeks a weighted estimator, some kind of weight smoothing should be done to make the estimator more stable. For simplicity and to minimize variance, we used one over the number of respondents in the domain $(1/n_r)$. This is approximately equivalent to using the average weight in the domain over the estimated domain size. The expected number of unique cases in the population for a given individual record can be estimated by its record multiplicity times the estimated probability of $(1-\hat{p})^{N_{non-collected}}$. If this number is greater than or equal to 1, we expect the record to be involved with at least one unique case in the population. In other words, the record is predicted as identifiable.

We define the record multiplicity limit by domain as the reciprocal of the estimated prediction probability of $(1-\hat{p})^{N_{non-collected}}$. Because of large variation in NHS sampling and response rates from domain to domain, the calculated limits vary from 2 to 431. A lower limit corresponds to higher NHS sampling and response rates. If we had a census in a domain and if everybody responded, then the limit would be 1; i.e., all observed unique cases in the sample are unique cases in the population. Obviously, a lower limit means a higher chance for an individual to be flagged as identifiable. Knowing this property, we deliberately set the limit to 1 for records that are from reserve or canvasser areas, where the NHS sampling fraction is 100%. We applied the same limit to individuals that were respondents of both the NHS and a postcensal survey. Consequently, we reduced more data for these individuals.

Note that the quality of this prediction approach is sensitive to the fraction of the domain population that is observed. If this fraction is too small, as an extreme case, the limit could actually be greater than the maximum attainable record multiplicity (i.e., the total number of tables produced for this domain). This implies that no record would be flagged for treatment, which might either be a sign of low prediction power or of low disclosure risk. It is safer in these cases to lower the limit to make sure a minimum number of records are treated (i.e., the ones with the highest record multiplicities).

5. Data treatment

Data treatment includes data reduction and data perturbation. To reduce the data of a record flagged as identifiable, we suppressed the values of some variables. We suppressed data record by record beginning with the record's worst variable in the analysis (i.e., the one with the highest variable multiplicity), then followed the decreasing order of variable multiplicities. The record multiplicity is reduced every time a value is suppressed. The suppression process goes on until the record multiplicity falls below the domain limit. We then applied residual suppression, which refers to the suppression of variables that are not indirect identifiers but are closely related to variables that were suppressed. We do this to protect suppression of indirect identifiers.

We suppressed data on the combined variables in two stages for the IPUMF in an effort to keep more useful data. In the first stage, we suppressed the less-essential or new components in 2011 of combined variables. We evaluated whether this suppression sufficiently reduced the record multiplicity by running the three-way tables again, with the reduced content of the combined variables. If it was not sufficient, we then suppressed the other components in the second stage. For example, for the combined variable POB_NEW, we suppressed the place of birth of parents in the first stage. Then, we re-applied the analysis of three-way tables on POB instead of POB_NEW. If the record was still at risk and POB still had a relatively high variable multiplicity, then we suppressed POB.

A specific variable may become less useful for analysis if it is suppressed for too many records. When this is the case, one may want to aggregate some categories of the variable. Aggregation reduces the number of suppressions because the aggregated categories contain more records, reducing the number of unique cases. We monitored the suppression rates of categories of variables. If the rate of a category was over 2%, we considered aggregation. For example, age groups could be aggregated from five-year to ten-year age groups. The broader categories mean less precision on individual data, which is a price to pay for lower suppression rates. A balance has to be achieved between having lower suppression rates and having more detailed categories. This process is called content revision. The redefinition of categories of variables is a consultation and negotiation process with subject-matter experts. It is an ongoing process, because aggregating the categories of one variable affects the suppression rates of other variables and their categories. Choosing which variable to aggregate for which subject is an art. In general, the more detailed the categories, the more analytical power the variable has. Subject-matter experts would prefer to avoid aggregating variables altogether. A balance has to be found among the subjects, and all parties should agree before any content revision. Content revision should also be consistent with the PUMF contents of previous production cycles. We proceeded by iteratively performing the analyses and suppressions and aggregating the variable values that were suppressed the most often, regardless of the subject, until all suppression rates were under 2%.

We applied data suppression or residual suppression to some numerical variables. On top of that, we subjected all numerical values to data perturbation, because a lot of the data came from administrative files. We applied random rounding to all the applicable values. Different variables use different rounding bases, which we determined by examining the characteristics and distributions of the variables' applicable values.

We treated extreme values by recoding and we identified those using thresholds by domain (e.g., geography by sex for total income). These domains are different from the confidentiality analysis domains. We defined the thresholds by studying the distribution of the weighted applicable values of the variables using the NHS data. Depending on the variable, the upper threshold could be the 99th, 98th or 90th percentile. In determining percentiles, we considered the nature of the variables and the input of subject-matter experts. We used top-coding to replace the value above the threshold with a weighted average of all the values of the NHS that are above the threshold. We then applied a residual top-coding method to variables that satisfied certain relationships with the top-coded variable. For example, after-tax total income is equal to total income minus income tax paid. If one of the three variables was top-coded, then we made sure at least one of the other two was top-coded as well. If not, then we top-coded the larger of the two. A lower threshold can be defined as a certain low percentile or simply as a value specified by subject-matter experts. We used the latter approach. Bottom-coding is done by replacing the value below the threshold with the threshold.

We also performed an outlier detection and treatment procedure at the end of the confidentiality analyses and the data treatments. It consisted of identifying very rare cases that would put an individual at risk of disclosure (e.g., non-single person younger than 20 years old) and of changing the values of some of the problematic variables. Very few records were treated this way.

6. Methodology differences of HPUMF

The HPUMF is more sensitive than the IPUMF in terms of confidentiality, since it contains hierarchical information on families and households. Below are the methodological differences in the development of the HPUMF. The methods that are not listed here but that were described in previous sections apply to the HPUMF as well.

- We used a smaller sampling fraction (1%).
- We used households as the sampling units in the second phase.
- We sorted data in the second phase by geography and by household, family and demographic characteristics.
- We protected very large households by keeping a maximum of seven people in the file.
- We included less geographical information in the file (16 identifiable geographies instead of 35 for the IPUMF). We defined the domains by crossing the geography and the household size.
- We used three-way tables at the household level. To do so, we used super-variables. We created them by sorting the individuals of the household in a certain order and by concatenating values for all members.
- We applied suppression to super-variable data (i.e., all values for the individuals in the household were suppressed at once).
- We did not suppress non-deterministically imputed values.
- In the content revision, we kept the suppression rates under 5% as much as possible.

7. Conclusion

A Public Use Microdata File can provide rich information, but privacy protection should always be the top priority. For the NHS PUMFs, we achieved this protection through data reduction and perturbation. We applied the data reduction method in sampling, in data suppression and in content revision. It was used mainly for categorical variables but was also applied to numerical variables. We used data perturbation for numerical variables. We applied data perturbation in weighting when we sought a self-weighted PUMF sample. The other applications of data perturbation were random rounding and recoding. The balance between more data and less risk should be considered throughout the development of the methodology, from sampling and weighting to confidentiality analysis and data treatment.

Acknowledgments

We thank Jean-René Boudreau and his team who developed the methodology for the 2006 Census long-form PUMFs, which was the starting point for this production cycle.

References

Särndal, C.-E., B. Swensson, and J. Wretman. (1992), *Model Assisted Survey Sampling*. Springer Series in Statistics. New York: Springer-Verlag.

Statistics Canada. (2010), *User Guide, Public Use Microdata File, Census of Canada, 2006, Individuals File*. Statistics Canada Catalogue no. 95M0028X.

Statistics Canada. (2011), *User Guide, Public Use Microdata File, Census of Canada, 2006, Hierarchical File*. Statistics Canada Catalogue no. 95M0029X.

Statistics Canada. (2014a), *User Guide, Public Use Microdata File, National Household Survey, 2011, Individuals File*. Statistics Canada Catalogue no. 99M0001X.

Statistics Canada. (2014b), *User Guide, Public Use Microdata File, National Household Survey, 2011, Hierarchical File*. Statistics Canada Catalogue no. 99M0002X.

Wolter, K. M. (1985), *Introduction to Variance Estimation*. Springer Series in Statistics. New York: Springer-Verlag.