Catalogue no. 99-137-X
ISBN 978-0-660-03896-4

A New Approach for the Development of a Public Use Microdata File for Canada's 2011 National Household Survey

by Jean-Louis Tambay, Ivan Carrillo-Garcia and Sri Kanagarajah

Release date: December 10, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Statistics Canada has agreed to produce a hierarchical PUMF that relies on perturbation instead of suppression to protect data confidentiality. This paper describes the creation of this new type of PUMF for the Agency.

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

1. Introduction

Statistics Canada has been producing anonymized Public Use Microdata Files (PUMFs) from the Census Program long form samples since the 1971 Census. For the 2011 National Household Survey (NHS), which replaced the long form, an individual (person level) and a hierarchical (household level) PUMF were produced using the same approach as previous censuses. Representative subsamples of 2.7% and 1% of the Canadian population, respectively, were taken from the 2011 NHS and global recoding and local suppression were used to protect data confidentiality. Additionally, in answer to requests from Integrated Public Use Microdata Series-International (IPUMS-International), and to meet the needs of users who prefer to work with unsuppressed data, Statistics Canada has agreed to also produce a third, hierarchical, PUMF that relies on perturbation instead of suppression to protect data confidentiality. This paper describes the creation of this new type of PUMF for the Agency. The next section gives background information on the project. Section 3 presents the approaches used for sample selection and disclosure control. The disclosure control strategy is centered on targeted data perturbation, whose application is covered in detail in Section 4. Concluding remarks are given in Section 5, which also provides summary results on the outcome.

2. Background

IPUMS-International disseminates integrated, confidentialized census microdata samples to researchers world-wide. As of 2013, 238 samples representing 74 countries were made available to more than 7,000 registered users, including several Canadian researchers (McCaa, *et al*., 2013). In response to requests from that organization, Statistics Canada has agreed to produce a detailed hierarchical PUMF from the 2011 NHS. Unlike current PUMFs produced by the Agency, the protection of data confidentiality would be achieved through data perturbation. The special PUMF would be provided to IPUMS-International to add to their collection after some manipulation on their part. IPUMS-International increases international comparability by adding metadata and content. For example, it creates standardized versions of variables and adds family relationships (Sobek and Kennedy, 2009).

Following discussions with IPUMS-International, Statistics Canada agreed that the special PUMF would include a set of 48 variables for a sample of about 2.7% of the population. Some variables, such as age, occupation and place of birth, would be very detailed, possibly at a cost of higher levels of perturbation.

3. Overall approach for creating the PUMF

The creation of the PUMF required finalizing the data content, selecting the 2.7% sample from the 2011 NHS, developing and applying a disclosure control strategy, and producing household weights for estimation and variance estimation.

3.1 Finalizing the data content

The data content was determined in consultation with IPUMS-International and subject-matter specialists. Variable categories were grouped in an attempt to balance confidentiality and analytical needs. The only geographical variables are Province, with the Territories combined, and a rural indicator, which was collapsed in PEI and the North. Instead of grouping, Age and income variables would be subjected to noise addition and top/bottom coding. For place of birth (POB), mother tongue (MTN), Occupation (OCC) and Citizenship, a minimum population size of 125,000 was used to determine the final categories. For OCC and POB, categories were grouped to respect the target. Smaller hard-to-collapse categories, for example Oceania, were left as is. For MTN and Citizenship, categories below the target were placed in a residual category. For Industry (IND), the 2-digit NAICS was generally followed. In the North, POB, MTN, derived Visible Minority (DVM), and non-Christian Religion (REL) categories with less than 400 people were put in a separate category. Those separate categories should have a negligible impact on analyses as they cover a very tiny portion of the Canadian population.

3.2 Selecting the public use microdata samples (PUMS) from the 2011 NHS

Sample selection was complicated by the presence of related PUMFs, which increases the risk of reidentification. As noted, two traditional 2011 NHS PUMFs were being produced. Additionally, post-censal surveys like the 2012 Aboriginal Peoples Survey (APS) and the 2012 Programme for the International Assessment of Adult Competencies (PIAAC), which selected Aboriginal and Immigrant households from the 2011 NHS, were also releasing PUMFs. To minimize the risk of reidentification, it was decided to avoid overlap with those PUMFs as much as possible. Overlap among the three 2011 NHS PUMFs was avoided by splitting the 2011 NHS into three portions, with 45/109th reserved for the present PUMF. By incorporating frame and design information from the post-censal surveys in the selection process, overlap with those surveys was nearly eliminated. The result was a 2.78% sample with minimum sample weight of 32 (shared by more than 85% of the households) and maximum weight of 242.22=10,900/45.

3.3 Developing and applying a disclosure control strategy

The disclosure control strategy is based on a mix of global recoding and targeted perturbation. As noted before, categorical variables were grouped to reduce the need for perturbation. The minimum threshold for categories for variables like POB and MTN was set at 125,000 after a threshold of 100,000 yielded too many problem cases. Further grouping was done in the North. Age and income variables were subjected to both noise addition and top/bottom coding. Age was top-coded at 85. Top codes for income variables were calculated by province (with the North treated separately), rural indicator and sex. For employment income and government transfers, 99th percentiles were used, whereas for other market income 98th percentiles were used. Values above each top code were replaced by their weighted mean. All income variables were bottom coded at -30,000 for women and for men in the Atlantic provinces and the North, and at -50,000 for men elsewhere. Perturbed income values were also rounded to base 100 except for nonzero values between -50 and +50, which were set to ±1.

Data perturbation efforts focussed on individuals and households with a high risk of disclosure. Candidates were identified through the application of rules and the use of disclosure risk measures.

For several years now Statistics Canada has been using disclosure risk measures as part of its strategy for creating PUMFs. For this project, it was decided to combine two of them: multiplicity and a Data Intrusion Simulation (DIS) measure. Both work from a set of *identifying variables (IVs)*, i.e., actual or derived PUMF variables that, in combination, can be used to identify individuals with unique values in the sample *as well as in the population* (e.g., a 68-year old male dentist in PEI). A microdata file can include a large number of IVs, but it does not make practical sense to use them simultaneously to identify individuals at risk – nearly everyone would be unique. Risk scenarios involving a few IVs at a time were thus created and results were obtained by subgroups created from widely or easily known characteristics (e.g., all combinations of three IVs, by province and sex). Subgroups are useful to include certain characteristics, like sex, in every risk scenario or when different subpopulations are at different levels of risk (e.g., respondents from smaller provinces are usually more at risk than those from larger ones).

The first risk measure, multiplicity, takes combinations of IVs, say three at a time within subgroup, and counts the number of times each unit is *sample unique* (i.e., the only respondent in his/her subgroup with a particular combination of IV values). The count, called multiplicity score, is related to the unit's identification risk since units that appear as unique in more tables are more likely to be identified as unique in the population. The multiplicity score is a heuristic concept. One problem with it is that it ignores features of the sample design that affect risk, such as the sample rates. Another problem is that it is difficult to come up with a theoretical threshold for a maximum acceptable score unless one is dealing with the simplest of sample designs (Bernoulli sampling).

The other risk measure is based on the Data Intrusion Simulation, a method of estimating the probability that, for a given scenario (set of IVs), a hacker who matched an arbitrary unit in the population against a sample unique on the PUMF is correct. For Bernoulli and Poisson sampling this probability is estimated from the number of sample uniques and pairs, and the average weight of units in pairs (Skinner and Elliot (2002), Skinner and Carter (2003)). The probability, calculated for any subgroup and scenario, can be assigned to every sample unique therein. This assignment can generate some peculiarities. For example the estimated probability for, say, a dentist who is sample unique may be affected by whether civil and electrical engineers are placed in the same or different categories. The estimated probability is 1 when there are sample uniques but no doubles for a given scenario. Its variance can be quite high. However, the simplicity and theoretical foundation of DIS makes it a very attractive tool for comparing strategies, like the impact of providing different levels of geographical detail on the overall risk.

The two measures were combined into a single unit-level risk measure as follows. Given a set of IVs, for n = 1, 2, and 3, all possible n-way tables by subgroup were generated, and the DIS probability for each table was assigned to each of its sample uniques. Rather than count up the number of times that a unit was unique, the five worst (highest) probabilities for that unit were taken. The probability was 0 for tables where the unit was not unique. The risk measure used, called $DIS_{(5)}$ here, *is the probability that the unit does not get correctly matched in any of the five n-way tables where it is most likely to be matched* (treating tables as if they were independent). If $DIS_{[i]}$ represents the i^{th} highest risk for a unit, then $DIS_{(5)} = 1 - \prod_{i=1}^{5}(1 - DIS_{[i]})$. Units with $DIS_{(5)}$ above our threshold were considered to be at risk.

The DIS probability estimates assume that the population is unclustered. The hierarchical nature of the PUMF necessitated some adaptation. Household members can share characteristics like place of birth, level of education and religion that will affect the table counts of uniques and pairs. In calculating person level risk, households were only allowed to contribute 0 or 1 unit to each cell. For household level risk the approach taken was to generate household level IVs and subgroups. Household IVs used in one-couple households include household type (e.g., three-generation household), the places of birth present, the highest level of education, the joint occupations of both spouses, the age/sex distribution of their children (counts for 13 age-sex groups collated into a single IV). Some household IVs, like the previous one, could have hundreds of categories.

Once units at risk were identified it was necessary to determine which variable(s) should be perturbed. The target variable was often determined using the risk measure again. We generated an individual $DIS_{(IV)}$ measure for each IV. $DIS_{(IV)}$ was similar to $DIS_{(5)}$ except that it was based on the five worst tables that did not include the specified IV. If a unit's $DIS_{(5)}$ was above our threshold, but some of its $DIS_{(IV)}$ were below, then those IVs were preferred candidates for perturbation. The choice of IV depended on factors such as how easy it was to perturb the IV and whether a particular IV was already perturbed enough times when it was the only choice (which was the case for OCC). When none of the $DIS_{(IV)}$ went below the thresholds, the IVs with the lowest $DIS_{(IV)}$ may have been perturbed, or more than one IV was perturbed. Sometimes, a different IV was used just because it was a good candidate for masking. For example, changing someone's DVM would make it much less likely for them to be identified through spontaneous recognition. For a few thousand households, both $DIS_{(5)}$ and the $DIS_{(IV)}$ never went below 1. Many of these so-called *unresolvable* households underwent more drastic perturbation measures such as changing province, adding, removing or swapping members, swapping members' education-employment histories, etc.

Candidates for perturbation were also identified by applying deterministic rules and using estimated population size thresholds. This was done particularly for ethno-cultural variables like POB, MTN, DVM and religion. Rules also targeted large households. Household size was capped at 11. Larger households were either split in two or had a few members removed. Large household below 11 were also subjected to more drastic perturbation. Section 4 gives more detail about some of the other deterministic and threshold rules used. Although for reasons of confidentiality and brevity not all methods used are presented, the section provides a good overview of the types of strategies used.

3.4 Producing household weights for estimation and variance estimation

After the perturbation was done, sample weights were calibrated so that PUMF estimates added up to Census population counts for 33 post-strata. Post-strata were generated by crossing province/the North, and three household types: households with 2 or more Aboriginal members, other households with 6 or more members and other households.

The PUMF sample is close to a self-weighting sample of households, with 85% of the weights equal to 32. As was the case for the traditional 2011 NHS PUMFs, the Random Group Method (Wolter, 2007) was used to allow users to generate variance estimates. With this method, sample households are randomly distributed among 8 replicates. Normally, following this step, each household has a replicate weight that is 8 times larger than its original weight, if it was selected for that replicate, and zero otherwise. The calibration step is then repeated for each replicate, leading to a set of 8 calibrated replicate weights for each household, which are equal to zero 7 times out of 8.

One problem with zero weights is that they can yield estimates of zero, which could lead to problems for certain types of analyses; a solution is proposed by Rao and Shao (1999). To avoid having replicate weights equal to zero, rather than using the replicate weights as is, each replicate weight was replaced by the simple average of the original weight and the replicate weight. This yields a weight that is one-half of the original weight when the household is not in a particular replicate. It is those weights that were calibrated. The result is a set of 8 nonzero replicate weights for

each household that can be used to generate replicate estimates for variance estimation purposes. More information on the estimation of variance using the (calibrated) replicate weights is provided in *2011 National Household Survey Special PUMF* (Statistics Canada, 2015).

4. Application of the data perturbation strategy

This section provides more information on the application of the data perturbation strategy, starting with the application of the risk measure.

4.1 Application of the disclosure risk measures

DIS risk measures were generated for both individuals and households. Disclosure risks for individuals were calculated only for members aged over 15 since little information was provided about younger members other than age, sex and ethno-cultural characteristics. Younger members would be covered by household level analyses and by the additional treatment of ethno-cultural variables. The risk analysis for individuals was done by subgroup defined by province, rural indicator, and sex. Within each subgroup, scenarios involving 16 IVs taken one, two and three at a time were used to calculate the $DIS_{(5)}$. IVs used were not necessarily variables as present on the PUMF. Quantitative variables like age and income were grouped into categories. Other variables were merged for convenience (e.g., Aboriginal identity was combined with DVM and the English/French mother tongue variables were combined with MTN). While the PUMF does not directly identify couples, some couples could be formed using the variables Relation to Person 1 (R2P1) and Marital Status. This allowed a distinction to be made between same sex and opposite sex unions in the IV marital status. Less than 1% of the approximately 770,000 respondents aged over 15 had a $DIS_{(5)}$ value above our risk threshold.

Using couples as formed above, household level risks were generated for four types of households: one-person households (around 98,000 households), one-couple households (216,000), multi-couple households (4,000) and other (no-couple) households (53,000). A set of IVs and subgroups was created for each household type. For one-couple households subgroups were defined by province, rural indicator, sexes of the couple (MM, MF, FF), and a variable called household class. Household class had categories such as couple only, with children (only), with one parent, with grandchildren, with a parent and children, with children and grandchildren, and various types of multi-generational households with one couple. The presence of unrelated members generated more household classes. There were 24 IVs including household size (slightly grouped), household income group, the couple's joint education, occupation or ethno-cultural characteristics, combined characteristics of their children or other members, dwelling characteristics, etc. Some variables, like the age-sex distribution of children, had hundreds of categories. As expected, the risk was much, much higher when working at the household level. Nearly 40% of the one-couple households had $DIS_{(5)}$ above our threshold, including several thousand with an estimated risk of 100%. For a large proportion of the households at risk, removing the IV for the joint occupation of the spouses brought the risk value $DIS_{(IV)}$ below our threshold.

Only about 4% of one-person households were above our risk threshold. The risk analysis for those households was done for 18 IVs, with subgroups formed by crossing province, rural indicator, and sex.

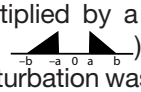
Results for the third largest group, no-couple households, were closer to those for one-couple households, with over 30% of household at risk. For these households 21 IVs were used and subgroups based on province, rural indicator, and household class. Household classes were lone-parent households, lone-grandparent households, three-generation households, other lone parents living with relatives, lone parents living with non relatives, other households of related individuals only, households of unrelated individuals only, other households with children, and other households. The IVs were created in a fashion similar to one-couple households.

The analysis for multi-couple households used 22 IVs and subgroups defined by region (Atlantic, Quebec, Ontario, Manitoba+Saskatchewan, Alberta, B.C., and the North) and household class, which could take up to 48 different values. The combination of small sample and multi-valued IVs and subgroups meant that all the households exceeded our risk thresholds; and for over $4/5^{th}$ the risk was 100%. Based on earlier studies involving hierarchical PUMFs these results were not surprising.

Once persons and households at risk were identified the next step consisted of identifying the variable or variables to perturb, as described in Section 3.3. For multi-member IVs, such as spouses' OCC, it was also necessary to choose which member(s) to perturb. Often OCC was singled out for perturbation. To reduce the extent of perturbation for this variable an alternate variable was selected whenever possible. Also, when the risk was not much higher than our threshold, perturbation was not carried out on a 100% basis. Perturbation was also avoided for some other variables in this step. This was the case for POB, MTN and DVM. These variables are extremely complicated to perturb because of the multiple relationships that exist between them (and with other variables like citizenship and religion) and between household members. Moreover, these variables were already subject to perturbation as a result of the ethno-cultural analyses (see Section 4.3).

Advantage was taken from the fact that variables such as income and occupation were imputed much more often than others on the 2011 NHS (around 10% for OCC). When these variables were imputed on the 2011 NHS, their values were treated as if they had already been perturbed, which was often enough to designate the unit as protected. (It may have been useful to include the fact that the values were imputed when calculating the risk, but there was not an easy way to do that.)

4.2 Perturbation of units at risk

Four types of perturbation were carried out. The simplest was the addition of random noise to quantitative variables like Age and income. When an income variable was selected for perturbation, its value was usually multiplied by a factor $1+\varepsilon$, where ε was a random noise that followed the split triangular distribution (shaped like 

). The perturbation of age was done in a similar way except for individuals aged 15 to 33, whose perturbation was made to better reflect their age distribution by marital status and highest degree completed (HCDD). Instead of perturbing those ages in each direction 50% of the time, the perturbation probabilities followed the age distribution of people with the same marital status in neighbouring ages. So a married 18 year-old would be much more likely to have his age increased rather than decreased.

The second perturbation method was swapping, which had the advantage of preserving univariate statistics. Swapping was applied to individual variables or to sets of variables between persons, to persons between households, and to households between regions. When the number of units to be swapped allowed it, efforts were made to swap among similar units. This could be achieved in two ways. First, swapping could be done within swapping cells created by crossing variables such as region, sex, income or education level. Second, units within a swapping cell could be sorted by an ordinal variable like income or education, so that individuals with the lowest/highest values for that variable would tend be swapped together. Proper swapping required a fair amount of preparation. For example, to swap the OCC of individuals in such a way that swapping partners never shared the same OCC, a minimum requirement was that a swapping cell could not have more than 50% of its members with the same OCC. If necessary, swapping cells could be collapsed or merged, or individuals could be moved to/from neighbouring swapping cells until the requirement was met.

The third perturbation method was the application of a specific change to a randomly selected set of units. For example, among households for which it was decided to change the number of children, a specific number were randomly chosen to gain children and the rest would lose children. Finally, deterministic and random perturbations were carried out. Such perturbations were particularly used in the treatment of ethno-cultural variables, which is described in Section 4.3.

The variable with the highest number of perturbation swaps was OCC. To minimize the impact of OCC swapping on different types of analyses, two swapping approaches were used more or less the same number of times.

The first consisted of clustering occupations and swapping OCC values within clusters. Using generalized principal components (PRINQUAL procedure in SAS®) we formed 21 classes of OCCs that are similar in composition, i.e., their mix-up of individuals with respect to POB, DVM, religion, year of immigration, income group, HCDD, field of study, region, age and sex. OCC swapping was done within OCC class, HCDD, and sex. When necessary to ensure full matches, some collapsing or category "jumping" was done. The clustering was used as much as possible. However, about a quarter of individuals had occupations that were alone in their class. OCC swapping for these individuals was carried out within groups generated using sex, employment income, rural indicator, and HCDD.

The second swapping approach was similar to one used by the U.S. Census Bureau and Westat (Krenzke, Li, and Zayatz, 2013). We formed swapping cells of individuals using a cross-classification of relevant variables. The most important variable is what Krenzke *et al*. (2013) call the cluster or prediction group. This is a grouping of individuals so that the individuals in the same group have similar predicted probabilities of belonging to the 70 different OCCs. The probabilities of belonging to different OCCs were modelled based on covariates POB, MTN, DVM, religion, year of immigration, income group, HCDD, field of study, region, age, sex, full time/part time work status, and school attendance. The subjects were then classified into 70, 58, and 25 groups according to their predicted probabilities.

Finally, swapping cells were created using the cross-classification of prediction group (70, 58 or 25, depending on donor pool size), income group, either skill type or skill level, sex, region, and survey weight group. Swapping was carried out within swapping cells as much as possible, but some collapsing of the least important variables was allowed when necessary. For a small percentage of cases the HCDD of the swapping partners were far apart. We redid the swapping for those individuals controlling for sex, HCDD, income group and, when possible, the 25-level prediction group.

The use of clustering helped to maintain relationships between OCC and related variables on the PUMF. To better preserve specific relationships, some variables were swapped alongside OCC. This was not done always, as there was a trade-off between preserving a variable's relationship with OCC on the one hand, and preserving its relationships with all other variables – and the desire to minimize overall perturbation rates – on the other. Industry (IND) was swapped alongside OCC always, and Field of study, most of the time. HCDD was swapped with OCC when the swapping partners were not sufficiently close with respect to HCDD, which was more likely when HCDD was not used in the original swapping.

Swapping was also used, to a smaller extent, for dichotomous and nominal variables at risk. Instead of swapping, perturbation for ordinal variables such as HCDD usually consisted of replacing values by neighbouring categories. This was done in a balanced manner to maintain marginal distributions as much as possible.

Aside from the random perturbation of Age, controls were applied to unusual differences in the ages of spouses and in the ages of parents and their children. The population distribution was used to set top and bottom codes for differences in spouses' ages. Great differences were reduced by changing one or both spouses' ages. A similar treatment was done for unusual differences in parent-children ages.

The most severe perturbation methods were used on unresolvables and on households at risk because of ethno-cultural variables but for which we did not want to change such variables. Methods used included changing the sex of members, changing their "life stories" (essentially the occupation and education variables), swapping persons between households, adding/removing children, and swapping geography. Swapping was usually carried out between similar persons/households.

4.3 Treatment of ethno-cultural variables

Ethno-cultural variables, particularly POB, MTN, DVM and religion can be problematic because they are more visible and because there are strong relationships between those variables for individuals and within households. Rules based on population thresholds were applied to those variables to identify individuals and households at risk. Examples are: values that are rare for their province; rare combinations of variable values for an individual; households with more than two category values for a variable (other than the commonest categories like POB in Canada, English or French MTN, not a visible minority...); rare mixes of values for spouses or within households, etc. In treating these rare cases we aimed to perturb the fewest members and values possible, but also wanted to respect relationships between variables and members. When a characteristic's value was changed the change usually affected all relatives with that value.

POB was easiest to change because its relationship between members was the weakest. Its large number of categories also made it the riskiest ethno-cultural variable. It was difficult to change DVM or MTN without changing POB as well (unless the change was to make these variables more "compatible"). Although the total number of perturbations was reasonably small, the treatment of rare cases did have an impact on the data. Univariate frequencies were generally not affected by much, although the rarest categories were affected the most. For POB (with Canadian POBs combined), MTN, DVM and religion, if we exclude the categories "Unspecified – Person lives in Northern Canada", the net impact of perturbation was never above 3%, and it was above 1.2% only 5 times.

Treatments for multiple and/or rare combinations of values within a household did make households slightly more homogeneous (e.g., households with too many POBs had POBs dropped). Rare combinations of variable values usually had one variable/value replaced, preferably by another value present in the household. This may reduce the occurrence of unusual cases such as immigrants from Europe whose visible minority status is of Asian origin.

Changes to POB and MTN sometimes triggered changes to other variables such as religion, citizenship, POB of Father/Mother, home language and language at work. POB perturbation was usually carried out separately for persons born in Canada and elsewhere to avoid having to change year of immigration.

5. Conclusion

The special PUMF included 925,564 individuals from 370,192 households. Excluding the small noise added to Age and income variables, 40% of individuals had values perturbed – 45% for individuals aged over 15. About 15% of individuals had more than one variable perturbed. Around 74% of the households with more than one member had at least one member's values perturbed beyond the "small noise". Although some practices such as top/bottom coding and the treatment of ethno-cultural variables may have increased data homogeneity, efforts were made to carry out perturbations in ways that reduced their impact on existing relationships. This was done particularly for variables that were perturbed more heavily like OCC.

Univariate population estimates from the PUMF were compared to those from the 2011 NHS. Differences could be introduced during the subsampling, perturbation and/or calibration steps (although calibration generally improved results). Excluding variables "Hours worked" and "Weeks worked," there were about 450 answer categories on the PUMF, and for three-quarters of them the difference with the 2011 NHS was within 1.25%. Only 23 categories had a difference of over 3%. For three of them (age categories 79 and 84 and Field of study category "Other") the difference was over 5%. Bivariate relationships were affected more, especially among rarer characteristics. It was following Subject matter reviews of such relationships that improvements were made to the perturbation of age and OCC.

The creation of this PUMF using data perturbation techniques, a first for Statistics Canada, was in many ways a research development project. In the process, ways were devised to avoid overlap with other PUMFs, to adapt and apply risk measures for a multitude of personal and household characteristics, and to carry out perturbations for related characteristics. During this process many lessons were learned whose benefit will extend to future work on PUMF creation at Statistics Canada.

References

Krenzke, T., Li, J. and Zayatz, L. (2013). Balancing Use of Weights, Predictions, and Locality Effects in a Model-Assisted Constrained Hot Deck Approach for Perturbation. 2013 *Joint Statistical Meetings Proceedings of the Survey Research Methods Section*, 1598-1612.

McCaa, R., Muralidhar, K., Sarathy, R., Comerford, M. and Esteve, A. (2013). Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten percent household sample of the 2011 census of Ireland for the IPUMS-International database. *Paper presented at the Joint UNECE/ Eurostat work session on statistical data confidentiality*, Ottawa, October 28-30, 2013.

Rao, J.N.K. and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 2, 403-415.

Skinner, C.J., and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society*, Series B, 64, 855-867.

Skinner, C.J., and Carter, R.G. (2003). Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling. Survey Methodology, 29, 2, 177-180. Statistics Canada, Catalogue No. 12-001-XIE.

Sobek, M. and Kennedy, S. (2009). The Development of Family Interrelationship Variables for International Census Data. Minnesota Population Center Working Paper No. 2009-02.

Statistics Canada. (2015). 2011 *National Household Survey Special PUMF (S-PUMF)*.

Wolter, K. (2007). *Introduction to Variance Estimation, 2nd* ed. Springer, New York.