User Guide for the Canadian Housing Survey Public Use Microdata File, 2018

Centre for Income and Socioeconomic Well-being Statistics,
Statistics Canada

Date released: 24/02/2021
Version 1.0

Table of Contents

1. Introduction: the Canadian Housing Survey (CHS)

The Canadian Housing Survey (CHS) provides information on how Canadians feel about their housing and how housing affects them. Information is collected on core housing need, dwelling characteristics and housing tenure, perceptions on economic hardship from housing costs, dwelling and neighbourhood satisfaction, perceptions on neighbourhood issues and safety, housing moves and intentions to move, volunteering, community engagement, life satisfaction, community satisfaction, dwelling adaptations to improve accessibility, self-assessed health, experience with homelessness, and sociodemographic characteristics.

The 2018 CHS was carried out in all 10 provinces and the 3 territories. The reference period was November 2018 to March 2019.

This guide provides information for users of data from the 2018 CHS Public-Use Microdata File (PUMF). The CHS PUMF is an anonymized microdata files that contains only a subset of variables that are available in the confidential CHS microdata files. Various techniques have been employed to protect CHS respondents against the risk of disclosure. For more information, see section 5.

This guide includes details on the survey methodology, data quality and key concepts of the survey. It also includes sections on the use, confidentiality and considerations for analysis in using the public use microdata file.

2. Survey methodology

The following sections outline 2018 CHS survey methodology[1].

2.1 Target population

The target population of the 2018 CHS is the population of Canada's 10 provinces and 3 territories, excluding residents of institutions, members of the Canadian Forces living in military camps and people living on Indian reserves.

People living in other types of collective dwellings are also excluded from the survey:

- People living in residences for dependent seniors; and
- People living permanently in school residences, work camps, etc.; and
- Members of religious and other communal colonies.

[1] For more details, see https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=793713.

For operational reasons, people living in some small remote areas in the territories where collection costs would be exorbitant are excluded from the survey. However, these people are included in the population estimates to which the CHS estimates are adjusted.

The survey covers approximately 98% of the population in the 10 provinces. The coverage of the territories was about 92% in Yukon, 99% in N.W.T. and 93% in Nunavut.

The CHS data for the Northwest Territories were obtained through a partnership with the NWT Bureau of Statistics. Instead of conducting the CHS in the Northwest Territories, data were obtained from the 2019 NWT Community Survey, which collects housing information similar to that collected by the CHS.

2.2 Sample

The 2018 CHS used the Dwelling Universe File as a frame. Administrative data on Social and Affordable Housing (SAH) were used to classify dwellings into strata in the frame. The frame was stratified into geographic areas of interest based on census subdivision (CSD) boundaries. Each geographic stratum was divided into two groups: SAH dwellings and all other (non-SAH) dwellings. Sub-strata were used to more efficiently sample rented and owned dwellings within the non-SAH strata, which are equally represented in the sample but not in the population.

For the provinces and the census agglomeration (CA) of Whitehorse, a systematic random sample of dwellings was selected independently within each stratum after sorting by household income, which was obtained from administrative files.

To reduce collection costs and the travel time for interviewers, the CHS sample in the territories (excluding Whitehorse) was selected using a two-stage design to allow for data collection by in-person interviews. First, towns (defined by CSD boundaries) were selected. Second, dwellings were selected within each of the selected towns to create the sample. Dwelling selection within towns was similar to dwelling selection in the provinces and based on a stratified, systematic random design.

For the purposes of the PUMF, to reduce the risk of disclosure, some of the original CHS geographic strata were grouped. For more information, see section 6.

2.3 Data collection

The 2018 CHS was conducted from November 2018 to March 2019. Data were collected using three collection modes: self-response Electronic Questionnaire (rEQ), Computer Assisted Telephone Interviewing (CATI) and Computer-Assisted Personal Interviewing (CAPI). CATI and rEQ were used exclusively in the 10 provinces and in the city of Whitehorse. CAPI was used in Nunavut and Yukon (excluding Whitehorse).

In the Northwest Territories, data were collected through the 2019 NWT Community Survey using CAPI and self-response. This collaboration between the CHS and the 2019 NWT Community Survey allowed for data to be collected on a wider range of topics and for a larger sample size in the Northwest Territories.

Whenever possible, the survey was completed by the household member with the most knowledge of the household's housing situation. In all cases, this person was aged 15 years or older. Proxy response was accepted for questions about other household members. This allowed one household member to answer questions on behalf of any or all other household members, provided they were willing and able to do so.

2.4 Data processing and quality control

The validation process consists of several activities to assess the quality of CHS data at specific levels of geography. As part of this evaluation, population and dwelling counts are compared with census data. The data quality measures used include response rates, invalid responses, and a comparison of data before and after imputation.

Tabulations for the 2018 CHS were produced and compared with corresponding data from the 2016 Census, other surveys and administrative data sources. Detailed cross-tabulations were also checked for consistency and accuracy.

Respondent data from the Northwest Territories were integrated into the CHS datasets.

2.5 Imputation of missing data

Individuals from a household were considered respondents if they met certain criteria after providing demographic information about themselves. A household was considered a respondent if at least one person in the household met the criteria. Any missing key data—either person-level data for individuals within the responding household, or data for household-level variables—were imputed.

Not all variables in the CHS database were imputed for non-response. Variables that were not imputed may contain fields with reserved code called "Not stated".

2.6 Weighting

The estimation of population characteristics from a survey is based on the premise that, in addition to itself, each sampled unit represents a certain number of non-sampled units within the population. A survey design weight is determined for each sampled unit based on its probability of selection to indicate the number of units within the population that the unit represents. Three types of adjustments are then applied to the survey design weights: non-response adjustment, calibration and adjustment for influential values.

For the PUMF, the weights of some households were further adjusted to reduce the risk of disclosure. Then all weights were recalibrated to ensure they matched population totals.

3. Data accuracy and quality

3.1 Sampling errors

Sampling errors arise when a population characteristic is estimated based on a portion or sample of the population rather than the population as a whole. Sampling error refers to the difference between the estimate derived from a sample survey and the “true” value that would be obtained if a census of the entire population were taken under the same conditions. There are no sampling errors in a census because the calculations are based on the entire population. The sample design and size, as well as the variability of the population characteristics measured by the survey, all affect the magnitude of the sampling error.

3.2 Measures of sampling error

There are several related measures of sampling error, and each one can be used to calculate the others using simple mathematical operations. The purpose of these measures is to estimate the degree of variation introduced in estimates by selecting one particular sample over another of the same size and design. Examples of commonly used measures are:

- Standard Error (SE), square root of the estimated variance of the estimate Y;
- Coefficient of Variation (CV), which is the calculated SE of an estimate Y as a percentage of the estimate Y: 100% × SE/Y;

- Confidence Interval (CI) for the estimate Y. This is the recommended measure to assess the variability (hence the quality) for proportions.

Because of the adjustments applied to the CHS weights for non-response and calibration, there is no simple formula that can be used to calculate variance estimates. Therefore, the bootstrap method for variance estimation—a pseudo-replication technique—is used to calculate the variability and report the quality of CHS estimates (ex: CIs are published as quality indicators for weighted frequencies and proportions).

Bootstrap weights are provided on the CHS master files but not on the PUMF. For more information and implications for analysis, see section 7.

3.3 Non-sampling errors

Non-sampling errors occur because certain factors make it difficult to obtain accurate responses or ensure that responses retain their accuracy throughout processing. Unlike sampling errors, non-sampling errors are not easily quantified. Four types of non-sampling error can be identified: coverage errors, response errors, non-response errors and processing errors.

a. Coverage errors

Coverage errors are omissions, erroneous additions, duplicates or unit classification errors in the survey frame and result from the incomplete listing and inadequate coverage of the population. Because they have an impact on each survey estimate, they are one of the most important types of error. They can introduce bias and the impact can vary for different subgroups of the population. These errors are often systematic and result more often in population undercoverage than in population over coverage. Survey weights are often calibrated to attempt to correct coverage errors. Slippage is a measure of survey coverage error and is defined as the percentage difference between control totals (post-censal population estimates) and weighted sample counts. In 2018, the CHS person-level slippage rate was 3.6%.

b. Response errors

Response errors can be attributed to many factors, such as faulty questionnaire design, misinterpretation of questions by interviewers or respondents, or errors in respondent reporting. These errors may be random, but can result in systematic bias if they are not. Great effort was made in the CHS to reduce response errors by implementing certain measures, including reviewing and testing the questionnaire using cognitive methods; using highly skilled and well-trained interviewers; and supervising interviewers to detect misinterpretation of instructions or problems with the questionnaire design. Response errors can also result from respondents—knowingly or not—providing inaccurate responses.

c. Non-response errors

Non-response errors result from a failure to collect complete information on all units in the selected sample. There are two kinds of non-response: total non-response and item non-response.

Non-response produces survey estimate errors in two ways. First, non-respondents and respondents often have different characteristics, which can result in bias. Second, non-response reduces the effective size of the sample, since fewer units than expected responded to the survey. As a result, the sampling variance increases and the precision of the estimates decreases.

Total non-response can arise for a variety of reasons—for example, if an interviewer is unable to contact a household, or no one in the household is able or willing to participate in the survey. Non-response adjustment of the survey weights for responding households is performed in the CHS to reduce the non-response bias.

Item non-response occurs when information is available for certain questions only because the respondent answered only a portion of the questionnaire. This occurs for a number of reasons. For example, a respondent may be unwilling or unable to answer the questions, the interviewer may fail to ask a question, or the wrong flow may have been followed through the questionnaire. Missing items are imputed to compensate for this partial non-response.

Non-response errors cannot be measured. However, these errors are generally significant when non-respondents differ greatly from respondents with respect to particular characteristics that are important determinants of survey results.

d. Processing errors

Processing errors are associated with activities conducted once the survey responses have been received, including all data handling activities after collection and before estimation. Processing errors can occur during any of the data processing stages, for example, during data entry, coding, editing, imputation, weighting or tabulation. Like all other errors, they can be either random—inflating the variance of the survey’s estimates—or systematic in nature—introducing bias. It is difficult to directly measure processing errors and determine their impact on data quality, especially since they are combined with other types of error (non-response, measurement and coverage). The use of a generalized processing system reduces potential processing errors. To minimize errors, diagnostic tests are carried out periodically to ensure that the expected results have been obtained.

3.4 Treatment of extreme values

For any sample, estimates can be affected disproportionately by the presence of extreme values from the population. For the CHS, a generalized processing system from Statistics Canada was used. This system can do recoding and apply edits. Editing processes are well defined with detailed specifications, for example, consistency and relationship edit specifications, and derived variable specifications. Validation is done after these processes are completed to verify any updates made to the data and ensure they meet the specifications. All values, including extreme values, are treated within this framework.

3.5 Response rates

Not all dwellings selected for the CHS were found to be in scope (occupied by an in-scope household) during collection. For example, some were occupied by temporary residents only, and some were occupied by households specifically excluded from the survey’s target population.

The overall response rate for the 2018 CHS was 50%. Response rates were calculated after counts of out-of-scope dwellings were removed.

4. Key CHS concepts and definitions

4.1 Tenure

‘Tenure’ refers to whether a household owns or rents their private dwelling. The private dwelling may be situated on rented or leased land or be part of a condominium. A household is considered to own their dwelling if some member of the household owns the dwelling even if it is not fully paid for, for example if there is a mortgage or some other claim on it. A household is considered to be renting their dwelling if no member of the household owns the dwelling, even if the dwelling is provided without cash rent or at a reduced rent, or if the dwelling is part of a cooperative.

4.2 Social and affordable housing (SAH)

Social and affordable housing refers to ‘non-market rental housing’. For this type housing, allocation and rent-setting mechanisms are not entirely dictated by the laws of supply and demand.

Due to the numerous and complex types of funding programs and agreements for SAH, households may not know that they are in SAH. The CHS collects information from the respondent about the presence of housing subsidies, the subsidy provider and the landlord to derive whether the housing is SAH.

4.3 Waitlist for social and affordable housing

Being on a ‘waitlist for social and affordable housing’ refers to the situation where the individuals seeking access to SAH have put their names on a waitlist. In the CHS, respondents are asked if anyone in the household is on a waitlist and how long they have been on the waitlist.

4.4 First-time homebuyers

First-time homebuyer, five years refers to individuals that purchased a home to live in less than five years before the reference date, and did not own a home before the purchase. Households are classified as first-time homebuyers if the purchaser and, where applicable, the cohabitating married spouse or common-law partner, are both first-time homebuyers at the time of the purchase.

4.5 Dwelling condition

‘Dwelling condition’ refers to whether the dwelling is in need of repairs. This does not include desirable remodelling or additions. Respondents classify their dwelling into one of three groups: needing regular maintenance only, needing minor repairs and needing major repairs.

Dwellings in need of major repairs are used as an indicator of inadequate housing by housing organizations, including Canadian Mortgage and Housing Corporation (CMHC). Major repairs include those to the dwelling structure or the major systems of the dwelling (heating, plumbing and electrical). The CHS questionnaire provided the following examples where ‘major repairs’ are needed: defective plumbing or electrical wiring, structural repairs to walls, floors or ceilings, etc.

4.6 Housing suitability

‘Housing suitability’ refers to whether a private household is living in suitable accommodations according to the National Occupancy Standard (NOS) and is measured by whether the dwelling has enough bedrooms for the size and composition of the household (such as age, sex, and relationships between household members).

For households in N.W.T., because data is integrated from the 2019 NWT Community Survey, gender is used in lieu of sex in the above rules for the NOS.

4.7 Housing affordability

The shelter-cost-to-income ratio, a measure of housing affordability, refers to the proportion of before-tax household income that is spent on shelter costs. Shelter costs for owner households include, where applicable, mortgage payments, property taxes and condominium fees, along with the costs of electricity, heat, water and other municipal services. For renter households, shelter costs include, where applicable, the rent and the costs of electricity, heat, water and other municipal services. Housing is unaffordable if the occupants of the dwelling paid 30% or more of their before-tax household income in shelter costs.

4.8 Core housing need

Core housing need is derived in two stages. The first identifies whether the household is living in a dwelling considered unsuitable, inadequate or unaffordable. Suitable housing has enough bedrooms for the size and composition of the household’s residents according to NOS requirements. Housing is adequate housing when its residents report that no major repairs are required and affordable when shelter costs are less than or equal to 30% of a household’s total before-tax income.

The second stage establishes whether the household has affordable access to suitable and adequate alternative housing by comparing the household’s total income to an income threshold based on local housing costs. Only those households that cannot afford alternative housing would be considered in core housing need. The income thresholds are derived at the census subdivision level by the CMHC. Income thresholds are derived from market shelter costs and are specific to the community in which a household lives.

Only private, non-farm, non-reserve and owner- or renter-households with incomes greater than zero and shelter-cost-to-income ratios of less than 100% are assessed for ‘core housing need’. Households not assessed for core housing need are excluded from the calculation of the core housing need rate.

Non-family households where the reference person is aged 15 to 29 and attending school are not considered to be in ‘core housing need’ regardless of their housing circumstances. Attending school is considered a transitional phase and the low incomes earned by student households are viewed as temporary.

4.9 Reasons for moving of the reference person

‘Reasons for moving’ refers to the reference person’s expressed reasons for moving dwellings. Respondents are asked for all reasons relating to their most recent housing move and the reason for the next intended move. Contextual information is also collected on when the last move occurred and when the next intended move is planned. Lastly, information is also gathered on whether or not household members are moving together as a unit to provide a more complete picture of people’s housing trajectories.

‘Reasons for moving’ is an important concept because it is at the heart of the issue of whether Canadians have housing that meets their needs and wants. Relocating is one way households address their housing issues or unmet needs. The reasons for moving can inform housing policies designed to address housing needs.

4.10 Dwelling satisfaction of the reference person

‘Dwelling satisfaction’ refers to the reference person’s subjective assessment of their satisfaction with their dwelling. Respondents are asked to rate their overall satisfaction on a five-point scale: “very satisfied”, “satisfied”, “neither satisfied or dissatisfied”, “dissatisfied” and “very dissatisfied”.

Dwelling satisfaction of the reference person is an important indicator because reference persons play an important role in housing decisions. Their perspectives on their dwelling can influence decisions to relocate or renovate, which can affect housing markets. Furthermore, integration of data on perceptions with traditional housing indicators—like core housing need—provides more information to measure whether housing needs are met.

Note: the PUMF has a four-point scale for dwelling satisfaction—the last two original categories “dissatisfied” and “very dissatisfied” have been grouped into a single category “dissatisfied or very dissatisfied”.

4.11 Neighbourhood satisfaction of the reference person

‘Neighbourhood satisfaction’ refers to the reference person’s subjective assessment of their satisfaction with their neighbourhood. Respondents are asked to rate their overall satisfaction on a five-point scale: “very satisfied”, “satisfied”, “neither satisfied or dissatisfied”, “dissatisfied” and “very dissatisfied”. The neighbourhood refers to the area surrounding the home.

As with dwelling satisfaction, neighbourhood satisfaction is an important indicator because it is related to housing decisions and housing need. Moreover, neighbourhoods and people’s perception of their neighbourhood are linked to concepts such as social inclusion. Indicators on neighbourhood satisfaction can inform policies on creating inclusive communities.

For households in N.W.T., because data is integrated from the NCS, the neighbourhood satisfaction refers to satisfaction with the community.

Note: the PUMF has a four-point scale for neighbourhood satisfaction—the last two original categories “dissatisfied” and “very dissatisfied” have been grouped into a single category “dissatisfied or very dissatisfied”.

5. The CHS public use microdata file (PUMF)

The CHS questionnaire contains information about households. The reference person (aged 15 years and older) responds for each household member. The 2018 CHS questionnaire is available online on the Statistics Canada website[2].

The CHS public use microdata file (PUMF) is an anonymized microdata file that contains only a subset of variables that are available on the CHS master files. While the CHS master files contain both person-level information and household-level information, the data contained in the CHS PUMF have been manipulated so that all the information is at the household level only.

[2] https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=getInstrumentList&Item_Id=1197668&UL=1V&

More specifically, this means that PUMF variables were derived using information about:

- the household in general (ex: core housing need)
- the person who answered the questionnaire (reference person, ex: their age)
- at least one member of the household (ex: highest level of education for the household)
- at least one member of the household who is related to the reference person (this would exclude roommates, lodgers, etc.; ex: visible minority household)

The CHS PUMF file contains 61,764 records. Each record has 132 variables. These variables can be grouped into the following broad topics.

Housing

- Housing situation (7 variables)
- Shelter costs and mortgages for owners (2 variables)
- Shelter costs and subsidy for renters (5 variables)
- Waitlist for social and affordable housing (2 variables)
- Previous accommodations (18 variables)
- Intentions to move (3 variables)

Dwelling

- Dwelling characteristics and tenure (4 variables)
- Dwelling satisfaction (12 variables)
- Dwelling issues (4 variables)
- Dwelling accessibility adaptation (5 variables)

Neighbourhood and community

- Neighbourhood services (19 variables)
- Neighbourhood satisfaction (1 variable)
- Neighbourhood issues (9 variables)
- Neighbourhood safety and crime (4 variables)
- Community satisfaction (2 variables)
- Civic engagement (4 variables)
- Volunteering (2 variables)

Health and well-being

- General health (2 variables)
- Life satisfaction (2 variables)
- Economic hardship (6 variables)

Household

- Household composition (7 variables)
- Household visible minority status (1 variable)
- Household Indigenous status (1 variable)
- Household highest level of education (1 variable)
- Household employment (1 variable)
- Household income (1 variable)
- Gender of the reference person (1 variable)

Other

- Geography (3 variables)
- Unique household identifier (1 variable)
- Version date (1 variable)
- Household weight (1 variable)

Additional information and a complete list of PUMF variables are available in the PUMF data dictionary.

6. Confidentiality of the public use microdata file

The production of a PUMF includes many safeguards to prevent the identification of any one household. Various techniques can be employed to protect respondents against the risk of disclosure. Some examples are included below.

- Data reduction involves limiting the amount of identifying information on the PUMF. Techniques include removing direct identifiers, sub-sampling, reducing the level of detail, grouping categories and suppressing data values for specific records.
- Data perturbation involves applying protective measures by coarsening or perturbing the data to hamper re-identification attempts. The addition of noise, random rounding and data swapping are common examples of perturbation techniques.
- Quantitative variables with very large positive or negative values are usually rare or unique in the population. Such extreme values are often top- and/or bottom-coded, which involves replacing the top or bottom values in a manner that preserves some aspects of the distribution while reducing disclosure risk.

All of the above measures have been applied when creating the 2018 CHS PUMF. Key treatments applied to PUMF variables include:

- Reducing the level of geographic detail on the file: the original census metropolitan area (CMA) and census agglomeration (CA) groupings on the master files were modified for the PUMF. Some smaller CMAs were collapsed with each other (Saint John with Moncton in New Brunswick) or the province (Lethbridge in Alberta, St. John's in Newfoundland), CAs were collapsed with outside CMAs/CAs, and the three territories were grouped into a single category. As a result, Prince Edward Island and Newfoundland are presented at the provincial level only. For more information, refer to the PUMF data dictionary.

- Grouping categories: for example, answer categories for questions about satisfaction with the dwelling and with the neighbourhood were grouped into fewer categories on the PUMF, as was mentioned previously.
- Data perturbation: the age of the reference person has been perturbed to avoid disclosure.
- Top/bottom coding and rounding: quantitative variables (e.g., income, shelter costs, outstanding mortgage amount and time on the waitlist for SAH) were all treated in this manner.

Note also that all income and income-related variables (e.g., core housing need, shelter-cost-to-income ratio groups) were suppressed for the territories for the purposes of the PUMF. As with other treatments, this was done do protect respondent confidentiality.

Furthermore, since the three territories were collapsed into one category, any variables that were coded as valid skips for N.W.T. because the corresponding questions were not asked on the 2019 NWT Community Survey, were suppressed for Yukon and Nunavut, to avoid the identification of respondents in these two territories.

Protecting the confidentiality of survey respondents is the top priority in the creation of a PUMF.

7. Considerations for analysis

The public use microdata files differ in many ways from the survey's confidential microdata (master) files held by Statistics Canada due to measures taken to preserve the anonymity of survey respondents, outlined above. There will be differences between estimates produced using the CHS PUMF and the estimates available in the data tables published on the Statistics Canada website, produced using the CHS master files.

Consequently, there will be limitations on analyses that can be conducted using a PUMF. The treatment of PUMF variables to reduce disclosure risk (outlined in section 6) introduces discrepancies between PUMF estimates and master file estimates. In addition, it is not possible to obtain variance estimates and therefore to produce quality indicators for the CHS PUMF estimates, since bootstrap weights are not included for this CHS PUMF cycle.

Users are encouraged to use the PUMF for data exploration and preliminary analyses, and then to request access to the confidential microdata from the Research Data Centres (RDCs) to conduct final analyses using the 2018 CHS master files.

7.1 Consistency with CHS published estimates

The magnitude of differences between estimates from the PUMF and the master files depends primarily on two factors: the level of geography and the number of cross-tabulations (domains of interest). Higher levels of geography with few or no cross-tabulations should result in PUMF estimates that are similar to the published ones or those calculated from the master files. A small geography, even with no cross-tabulations, may result in PUMF estimates that are significantly different from the master file estimates for the variable of interest. Once the variable of interest is cross-tabulated by domains of interest, the differences may become significantly large. This happens because the impact of suppressions imposed to reduce disclosure risk often becomes more pronounced—results in more distortions—as the level of detail increases.

For example, for a large province like Ontario, the estimate of dwellings that are part of a condominium is 14.4% from the master file and 14.0% from the PUMF, which is close. In contrast, for a small province like New Brunswick, the same estimate is 3.0% from the master file and only 1.0% from the PUMF, which is quite a large difference. Looking at another small geography, Prince Edward Island, the estimate of households in core housing need is 8.6% from the master file and 6.4% from the PUMF. When cross-tabulated by tenure (owner/renter households), the master file estimate of owner households in core housing need is 6.4%, whereas the corresponding PUMF estimate is 4.6%.

Such discrepancies are examples of how the treatments of records and variables on the PUMF (subsampling, suppressions, etc.) distort the estimates obtained using confidential microdata. The trade-off between protecting confidentiality and achieving consistency with actual estimates is a reality of every PUMF file. When it is not possible to satisfy both objectives, protecting respondent confidentiality must always be the top priority.

7.2 Point estimates

Microdata users should be aware that the results of hypothesis tests (such as the *p* values accompanying *t* statistics or Pearson statistics), provided automatically by standard statistical software packages, will be incorrect for data provided by surveys with a complex survey design, such as the CHS. For complex surveys, variance estimation and calculations of quality indicators such as Standard Errors (SEs) or Confidence Intervals (CIs) are done using bootstrap weights. The PUMF for the 2018 CHS cycle does not contain bootstrap weights. Therefore, only (weighted) point estimates may be calculated.

Without SEs or CIs, users cannot assess properly whether PUMF estimates for are the same for two areas or domains of interest. For example, an estimate of 20% with a confidence interval (CI) of (18%, 22%) is much more precise—and thus reliable—than an estimate of 20% with a CI of (8%, 33%). Furthermore, users cannot conclude that an estimate of 12% with a CI of (9%, 16%) is different from the estimate of 20% with a CI of (8%, 33%).

For this reason, comparative analyses conducted using the PUMF are strongly discouraged because the conclusions may not be valid, and users may get different results when using the CHS master files.

7.3 Reserved codes

If the coverage of a variable does not extend to a certain population subgroup, then there are no valid values for that subgroup. The values (reserved codes) that do appear are in the form of 6, 96, 996, 999.6 and so on and indicate that the variable is not applicable (valid skip). The coverage of each variable on the file is referred to in the data dictionary as the “universe”.

For certain records, no valid response is available, although the question is applicable. On the CHS master file, this corresponds to the “Not Stated” category. These missing values appear with a reserved code such as 9 or 99. On the PUMF, any value suppressed for confidentiality reasons will also appear with a 9, 99, etc.

It is important to account for reserved codes in any analysis. For numeric variables, if the calculation of means or aggregates seems too high, users should check to ensure that reserved codes were excluded from the calculation. For categorical variables, users should if possible, assess the impact of missing values on the overall representativeness of the data (for example, by comparing PUMF estimates with those released officially). In many instances, it will be appropriate to exclude valid skips and not stated categories when looking at distributions of frequencies and proportions.

7.4 Guidelines for analysis

Given the limitations outlined above, the CHS PUMF should be used with caution. Here are some guidelines for the use of the PUMF.

Uses of the PUMF:

- To conduct exploratory data analysis in order to become familiar with the data and weights; to see what kind of information is available and at which geographic level; as well as to assess the feasibility of variables and domains of interest;
- To determine what type of analysis is appropriate (ex: regression, etc.);
- To obtain preliminary estimates, keeping in mind that estimates may change (as may the conclusion) when the same analysis is conducted using CHS master files.

To avoid the impression that calculated estimates are more precise than they actually are, it is recommended that users round PUMF estimates in the same way as published CHS estimates by:

- Rounding weighted frequencies to the nearest 100 (e.g., 12,146 becomes 12,100); and
- Rounding weighted percent to at most one decimal place (e.g., 17.53% becomes 17.5%).

Users should remember to remove non-applicable categories from analysis. These categories include “Valid Skip” and, most of the time, “Not Stated”.

Users should also keep in mind that collapsing of geography variables on the PUMF precludes analyses for the individual territories and at the national level where variables or domains of interest were suppressed for the territories on the PUMF. For example, 2018 CHS shows that 11.6% of Canadian households are in core housing need (CHN). The PUMF estimate is also 11.6%, but now it represents households in CHN in the provinces only, since CHN was suppressed for the territories on the PUMF.

Finally, users must always use the provided weight, PFWEIGHT, for all analyses and inference. Unweighted counts differ from weighted counts (for example, due to the oversampling of certain populations of interest), and conclusions based on unweighted data will likely differ from those obtained using weighted data.

Do not use the PUMF to calculate estimates for policy work or any other purpose requiring estimates with known quality. PUMF estimates cannot be assessed for quality and, therefore, cannot be assessed for fitness for use.

8. Documentation and references

8.1 PUMF products available

The following PUMF products are available:

- One household level PUMF file in the following format: text (.txt), SAS, SPSS, Strata.
- Data dictionary
- *“User Guide for the Canadian Housing Survey Public Use Microdata File, 2018”*

8.2 How to quote PUMF

Please use the Digital Object identifier associated with this product to quote.

DOI: https://doi.org/10.25318/46250001-eng

8.3 Related products and services

There are a number of pre-linked datasets available for the CHS in the Research Data Centre, including:

- Administrative tax data
  - Administrative Personal Income Masterfile (APIM) (2018)
  - T1 Family Files (T1FF) (annual 2008 to 2017)
  - T4 Tax Files (T4) (annual 2008 to 2017)
- Historical Addresses File (tax derived)
- Income Dispersion File (tax derived)
- Proximity Measure Database
- Social Inclusion Index

These datasets can be combined with the core CHS questionnaire (household and personal) data using direct household (MASTERID) and personal (ID_PERS) identifiers. Some datasets only cover a specific subpopulation and will not have a corresponding record for all CHS records, specifically:

- The T1FF and T4 datasets only have responses for persons linked to an administrative record

Users are advised to treat missing values that are present in the administrative data or are missing for CHS respondents as not in scope for the pre-linked dataset.

For clients with more specialized data needs, custom tabulations can be produced to their specifications on a cost-recovery basis under the terms of a contract (subject to confidentiality restrictions).

8.4 CHS documentation

Definitions, data sources and methods for this survey are available on Statistics Canada's official website. To view the CHS questionnaire, visit:

https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=793713.

8.5 CHS publications

Data tables, articles and analytical papers are available at the link below.

https://www150.statcan.gc.ca/n1/en/surveys/5269

9. Questions and comments

If you have any questions or comments about the microdata, you can get in touch with the **Contact Us** team:

Phone: 1-800-263-1136 or 514-283-8300
Email: STATCAN.infostats-infostats.STATCAN@canada.ca

Centre for Income and Socioeconomic Well-being Statistics Canada
Ottawa, Ontario
K1A 0T6