# *2012 Canadian Nature Survey* Public Use Microdata File

## Data User's Guide

Thank you for your interest in the results of the *2012 Canadian Nature Survey*.

You are encouraged to read the report on the results of the *2012 Canadian Nature Survey*, which can be downloaded free of charge from www.biodivcanada.ca.

In particular, please review the section entitled **Survey Methods** which provides detailed descriptions of the sampling approach, the numbers of respondents by geographic and demographic category, the data collection process, the weighting applied to the different samples, and advice on using the data for further analysis.

The *2012 Canadian Nature Survey* **questionnaire** is provided as a separate file and is reproduced in the above report in Appendix B: Survey Instrument.

### Applying weights to records in the dataset

Each record in the dataset contains **two (2)** analytic weights: one of these **must** be used when conducting data analysis.

Responses to the 2012 Canadian Nature Survey were weighted to ensure that survey estimates would be representative of the Canadian population. Two sets of weights were computed: one "address" weight for data analysis from the address-based probability sample and a second "combined" analysis weight to support analyses of Web panel and opt-in responses. Both the address-based and combined sample weights will estimate similar population totals and distributions with respect to the variables used in the calibration adjustments. The difference between the address-based sample weight and the combined sample weight is that the former is based on a probability sample and can be relied upon for estimates and analyses that seek to draw conclusions about the Canadian population with quantifiable precision using standard procedures. The combined weight provides representative estimates with precision that cannot be estimated using standard procedures. That is, variances from the address-based sample can be estimated, but not, given standard statistical methodologies, from the combined sample.

The two weights allow two general types of analysis to be conducted. It is recommended that the choice of weight be based on the user's goals:

- The address-based sample weight should be used for analyses that require confidence intervals (e.g., population estimates) or statistical testing;
- The combined sample weight can be used for analyses that do not require confidence intervals or statistical tests (which require variance estimates). These data provide a representative set of responses from Canadian adults whose statistical reliability cannot be quantified.

For more information, please see the section entitled Survey Methods in the report *2012 Canadian Nature Survey: Awareness, participation, and expenditures in nature-based recreation, conservation, and subsistence activities*.

### Address-based sample weight

The weighting approach for the address-based sample is made up of two components: a sampling weight, which adjusts for unequal selection probabilities due to the stratified design, and a calibration adjustment, which adjusts for survey non-response and calibrates the weights to known population totals.[1] The sampling component of the address-based weight is the

[1] Population counts by age (for persons age 18 and older) and gender were obtained from the 2011 Census of Canada. Statistics Canada, 2011 Census of Canada: Topic-based Tabulations: Age and Sex for the Population of Canada, Provinces, Territories, Census Divisions, Census Subdivisions and Dissemination Areas, 2011 Census. Catalogue number 98-311-XCB2011018. (Ottawa, Ont.: Statistics Canada, 2012), http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/tbt-tt/Ap-eng.cfm?LANG=E&APATH=3&DETAIL=0&DIM=0&FL=A&FREE=0&GC=0&GID=0&GK=0&GRP=1&PID=101998&PRID=0&PTYPE=101955&S=0&SHOWALL=0&SUB=0&Temporal=2011&THEME=88&VID=0&VNAMEE=&VNAMEF= [accessed June 20, 2013]; Population counts for urban and rural areas were obtained from Gordon Dewis, Statistics Canada, email communication, June 2013; For a more in-depth discussion on population centres and rural areas, see Statistics Canada, "From urban areas to population centres," February 7, 2011,

inverse of the probability of selecting a given residence from the Canada Post address file within each sampling stratum. The sampling component was then calibrated (or post-stratified) to known population totals. This adjustment was made along the following dimensions:

- Age (independently within each province and territory)
- Gender (independently within each province and territory)
- Urban/Rural (independently within each province and territory)
- Aboriginal/non-Aboriginal identity

The result of this adjustment is the final address-based sample weight.

**Combined (mixed-sample method) sample weight**

The weighting approach for the combined sample has two components: a sampling component based on an estimated probability of selection for each respondent, and a calibration adjustment, which adjusts the first component to known population totals. The unknown selection probability for individuals responding from the Web panel and opt-in samples is estimated by modelling their propensity of inclusion in these samples using three stages. First, the address selection probability of each case was estimated as if it were selected via the address-based sampling mechanism. This is the same as the address-based probability for respondents in the address-based sample.

Second, the probability of being selected for the Web panel sample[2] was estimated by the modelled probability of being included in the panel data. That is, the event of responding via the non-probability sample is taken as the event of being sampled for the non-probability sample. The modelling relies on logistic regression to predict propensity of inclusion based on geographic location and most survey responses. In provinces and territories where both an address-based sample and Web panel sample were used, logistic regression was used to combine respondents from the Web panel survey and respondents from the address-based sample who responded on the Web. Predictor variables included response and demographic data obtained from survey responses, as well as the stratification variables. Logistic regression yields a probability of being selected for the panel for each respondent (not just those who answered using the Web).

Finally, the third step accounted for the fact that there were now two probabilities of selection (address-based and panel propensity). The two probabilities were combined through a procedure known as cumulating cases.[3,4] This procedure calculates the probability of a respondent being selected for the address-based sample, the panel propensity sample, or both. The sampling component of the combined sample weight becomes the inverse of this estimated probability of selection. For the Northwest Territories and Nunavut community opt-in survey, a constant selection probability was assumed. To produce the combined analysis weight, the sampling component of the combined sample weight was then calibrated (or post-stratified) to known population totals, using the approach described above for the address-based sample weight for Web surveys.

---

http://www.statcan.gc.ca/subjects-sujets/standard-norme/sgc-cgt/notice-avis/sgc-cgt-06-eng.htm [accessed June 21, 2013]; Population counts by Aboriginal identity were obtained from Statistics Canada, Table: Aboriginal identity population by age groups, median age and sex, 2006 counts for both sexes, for Canada, provinces and territories, http://www12.statcan.ca/census-recensement/2006/dp-pd/hlt/97-558/pages/page.cfm?Lang=E&Geo=PR&Code=01&Table=1&Data=Count&Sex=1&Age=1&StartRec=1&Sort=2&Display=Page. All totals were adjusted to be consistent to 2011 data.

[2] The selection process for both inclusion into the general Web panel, and for participation in this survey, are proprietary processes employed by the third party Web survey vendors.

[3] Steven Pedlow, Kanru Xia and Michael Davern, "Dual-Frame Sample Sizes (RDD and Cell) for Future Minnesota Health Access Surveys" *Proceedings of the Survey Research Methods Section,* American Statistical Association (2010): 2279-2288, http://www.amstat.org/sections/srms/proceedings/y2010/Files/307156_57962.pdf.

[4] Colm O'Muircheartaigh and Steven Pedlow "Combining samples vs. cumulating cases: A comparison of two weighting strategies in NLSY97," *Proceedings of the Survey Research Methods Section,* American Statistical Association (2002): 2557-2562, http://www.amstat.org/sections/srms/proceedings/y2002/Files/JSM2002-001082.pdf.

**Data dictionary**

A data dictionary is provided as a separate file. The data dictionary defines shorthand codes used for fields and values that appear in the dataset and where applicable provides counts for the number of responses to different options within questions. Note that the counts are based on the combined address based and web samples.

**Comparison with previously published results**

Analysis of the data may produce different results compared with those published in the report *2012 Canadian Nature Survey: Awareness, participation, and expenditures* because of differences in methodology or because of slight differences in the data resulting from the disclosure control strategy applied to the data (see below).

**Disclosure control**

Disclosure control is a process applied to a dataset to assure participants' confidentiality is preserved. The following disclosure control methods have been implemented in the preparation of the public-use microdata file in order to ensure the privacy of survey participants:

- Postal codes have been removed for all records.
- Year of birth has been re-coded to five-year age groups for all records.
- Records from the Nunavut community sample have been removed (57 records). Please see section in the report entitled **Survey Methods** for information on the Nunavut community sample.
- Text-based responses to open-ended questions that invited participants to provide further information if they selected "Other, please specify" have been removed.
- In addition, 19 records from web-panelists were removed due to spurious responses.