Producing official statistics via voluntary surveys – the National Household Survey in Canada

Marc. Hamel*

Statistics Canada, Ottawa, Canada, marc.hamel@statcan.gc.ca

Abstract

Statistics Canada conducts over 350 business, social and institutional surveys a year. Of all social or household type surveys, only one is conducted on a mandatory basis, the Labour Force Survey. By their very nature, voluntary surveys will achieve lower rates of response and are exposed to higher risks of bias. For the 2011 Census of Population program, the detailed form was for the first time collected on a voluntary basis as the National Household Survey. The survey content was basically the same as that of previous Census detailed forms and covered various socio-demographic topics that are of high importance to a wide variety of stakeholders in Canada. Given that one of the key characteristics of the survey is to produce data for small regions and for subgroups of the population, collecting it on a voluntary basis introduced several challenges. Statistics Canada, based on its extensive experience with voluntary surveys, developed a number of processes and approaches to ensure the highest data quality possible. This paper will describe what these measures were for data collection, data processing and estimation. It will also provide a brief description of the quality assurance framework underlying the release strategy of the 2011 survey.

Key words – Voluntary surveys, data quality, household surveys, response bias

1. Introduction

The Census of Population in Canada has been conducted every five years since 1956. Since the 1971 Census, basic demographic content has been collected of the entire population, and the more detailed socio–demographic and economic content has been collected on a sample basis, initially on a 30% sample, and then on a 20% sample from 1981 to 2006. In more recent censuses, the detailed questionnaire included more than 60 questions, on topics such as basic demographics, family and households, ethnic diversity and immigration, incapacity, education and training, language, Aboriginal Peoples, labour force status, income, pension, housing conditions and commuting to work. Until 2011, both forms were collected on a mandatory basis.

Concerns related to privacy of respondents have existed for some time in Canada. These were initially related to the enumerator based collection approaches where enumerators tended to work in areas where they themselves resided. Respondents were concerned about sharing personal information with neighbors. Collection strategies were adapted overtime to reduce the reliance on local enumerators, and saw the introduction of an Internet option and of a mail back process to a central location in 2006. Concerns were also expressed related to how the data were processed, especially related with the contract with Lockheed Martin in 2006 and 2011. Finally increased concerns were expressed related to some of the content covered in the questionnaire. In reaction to these privacy concerns, the Federal Government instructed Statistics Canada to collect the detailed content on a voluntary basis for 2011. This was done

via the new National Household Survey. The survey included the same content that was tested for the Census detailed questionnaire.

2. Description of 2011 National Household Survey

The National Household Survey became the biggest voluntary household survey ever conducted by Statistics Canada. The sample included approximately 4.5 million private dwellings, 30% of all private dwellings in Canada. Reference day for the survey was May 10, 2011, the same reference day as the Census. Almost all Canadians were in scope for the survey. The only exclusions were residents of collective dwellings (old age home, prisons, hotels, work camps, etc) and some small groups such as foreign diplomats. Since this was the first time such a large survey was conducted on a voluntary basis by Statistics Canada, no specific response rate objective was set although the assumption was that overall response could be 50%.

3. Planning the National Household Survey

Collection operations for the NHS were developed to maximize response in every region, attempting to minimize non-response biases while efficiently using the collection resources available to the survey. It was anticipated at the start of collection that there may not be sufficient financial or human resources available to follow-up every non-response to the survey. Operations for the survey were designed to be integrated with the collection and operations of the Census of Population.

The survey used two questionnaires, one for the web approach and a paper version. The online solution for the Census was adapted to introduce the survey to sampled households as soon as they submitted their census questionnaire. Every sampled household for the survey which answered the census online in May was presented with their survey questionnaire online right away. Common demographic questions were prefilled with the information provided in the Census.

As this was the first time the survey was conducted, there was concern that respondents may confuse the census to also be voluntary. For this reason, the survey was mostly introduced to sampled households only once they completed the census. Sampled households responding to the Census by mail received their NHS questionnaire by mail in early June or simply received the visit of an enumerator during non-response follow-up operations. The same follow-up approach was used for those who were non-respondent to the survey when presented online in May.

4. Non-response follow-up operations

To maximize response to the survey in every area, collection operations were managed at the collection unit level. Collection units (CU) include approximately 300 dwellings on average. There were more than 45,000 CUs in Canada in 2011 and approximately 30,000 enumerators were hired to follow-up non-respondents for both the Census and the survey at the same time in June, July and part of August. Although follow-up operations for the census and the survey were integrated, priority was given to the census in most cases. Follow-up on the survey began in a given area once Census non-response follow-up was sufficiently advanced. This was done to avoid compromising the quality for the Census.

On July 14, a sub-sample of approximately 400,000 of the remaining 1.2 million non-respondents to the survey were selected for further follow-up. This approach was initiated to better use remaining human

and financial resources in non-response and to ensure an acceptable level of response to the survey for every dissemination area. The distribution of the sample was determined based on observed levels of non-response at the time of sub-sampling, and on a factor of heterogeneity of the population. The more heterogeneous was the population in a given area, the larger was the sub-sample relative to the level of response at the time. The heterogeneity of the population was determined based on information from the 2006 Census long-form and was calculated for areas of about 400 dwellings. As collection progressed, follow-up would stop once a certain level of response was attained at the CU level. This was controlled centrally using a dynamic model which considered the level of response at the time, the staff hours available in the area, and the budget remaining. As collection stopped in some CUs, effort and resources were redistributed to remaining CUs. In addition to enumerators in the field, available capacity in Statistics Canada call centers were also used for the follow-up operations. Close to 100,000 cases with telephone numbers were sent to the call centers.

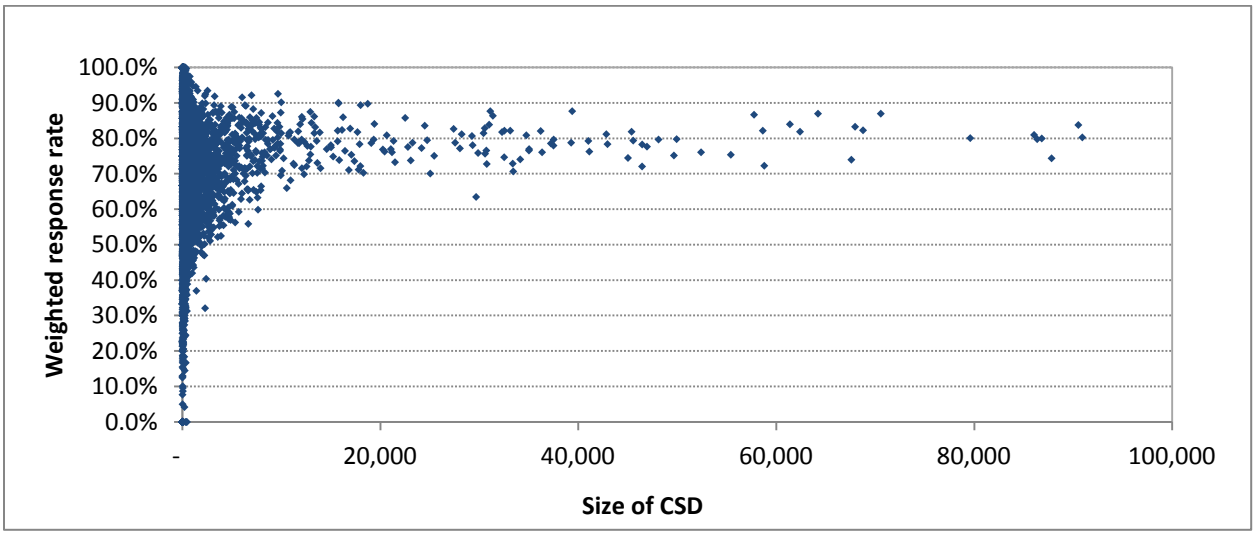
5. Response rates to the survey

The final response rate to the survey at the end of collection and after all returns were processed was 68.6%. This rate is comparable to other voluntary household survey conducted by Statistics Canada. A larger share of the response rate was provided by the internet returns at 43.5%.

Since a sub-sample of non-respondents was selected to complete collection, this is considered in calculating the response rate. Responding households selected in the sub-sample also represent the other non-responding households not selected in the sub-sample. The resulting weighted response rate for the survey is 77.2%. The unweighted response rates ranged from 60.4% to 71.9% among the 10 Canadians provinces. In the 3 northern territories, it ranged from 64.9% to 83.9%. For many communities in the territories, collection was conducted using a canvasser approach on a 100% sample.

As the objective of the survey is to produce data for small areas and small populations, it is important to consider response for lower levels of geography. One of the geographical units of interest for data users are Census Sub-Divisions which are the equivalent of municipalities and towns. Response at that level is less evenly distributed than at higher levels, ranging from 0% to 100%. The response level varies more as CSD are smaller.

Distribution of weighted response rate to the NHS by size of Census Sub-Division (CSD), CSDs of 100,000 occupied dwellings or less



Another element of the response to the survey is the response level to each item. Item response varies a lot for the NHS depending on the section of the questionnaire. Item response ranges from 96.7 % to 99.7% for the socio-demographic questions which are situated at the start of the questionnaire, but is lower for the items on education (89,4 % to 95,8 %) and items on labour, income and dwelling characteristics (80,7 % to 93,9 %). As with previous census detailed questionnaires, these items were at the later part of the questionnaire and usually require a bit more research on the part of respondents to provide a more accurate response. The higher non-response to these items was observed both on Internet and on paper returns.

6. Data Processing

Returns for the survey were processed with the same systems developed for the Census. Paper returns were captured using high performance optical reading and character recognition software. All returns went through the same editing steps and automated coding processes. These systems were verified to produce high quality outputs. All manual interventions that are part of data processing (e.g. keying from image, manual coding) were subject to sampled verification for accuracy.

7. Edit and imputation and weighting

In the editing phase, all items that do not have a value entered but that should been responded were identified. The missing values were imputed using a nearest neighbor approach. Replacing values were obtained from records that are similar to the ones with item non-response using factors of proximity that are based both on geography and key characteristics of both the donors and the receiving records.

The final responses were then weighted to represent the target population for the survey. A sampling weight was first assigned to each selected dwelling representing the inverse of the probability of selection. The weights were then adjusted to account for the selection of the sub-sample of non-respondents in July. In this phase, weights were further adjusted to account for dwellings which were still non-respondents at the end of collection. Weights of non-respondents in this phase were transferred to respondents using a

nearest neighbor approach similar to the one used in imputation. In the final phase, the weights were calibrated to census totals at the weighting area level. Weighting areas include on average 2,300 dwellings or 5,600 people. They are created by grouping several contiguous dissemination areas and need to be large enough to support the calibration. It was not always possible to respect the boundaries of census divisions and sub-division in forming the weighting areas. The NHS being an omnibus survey, the calibration was attempted on a total of approximately 60 variables. Control totals included age, sex, marital and common law status, dwelling type, household size, family composition, and language. Many calibration controls had to be relaxed during the process to preserve the overall quality of the detailed estimates.

Because weighting areas may include several Census Sub-Divisions, there sometimes exist differences between the Census and NHS estimates for common characteristics. The smaller the CSD, the higher is the risk that NHS estimates will be different than those of the Census for those characterictics.

8. Data Certification

Final results from the survey were compared to a number of external sources to verify their coherence and quality. These sources include the 2006 and 2014 Censuses, other Statistics Canada surveys, administrative sources such as federal tax data, the Indian Register, the longitudinal immigration file, etc. Most sources do not support certification of the results at low levels of geography. Comparisons were made at the national, provincial, territorial, and in some cases Census Metropolitan Area level. Results at lower levels were reviewed for overall integrity, focusing more on outliers and large changes compared to the 2006 Census.

9. Quality indicators for dissemination

The Census of Population uses a measure of global non-response (GNR) to determine what data is suitable for dissemination. This approach was also used for the census long-form prior to 2011. The measure combines the non-response at the household level with item non-response. The GNR is calculated for every region for which data are to be released.

It is well known that as non-response increases so is the risk of bias as non-respondents tend to have different characteristics than respondents. Statistics Canada conducted a bias analysis for NHS key estimates in relation to the GNR. From this analysis, an acceptable level of GNR for the NHS was established at 50%, meaning that results for standard data products disseminated on the web would be released for all areas with a GNR of 50% or less.

The bias indicators were produced from a linked file of the 2011 and 2006 censuses. Using family name, address and date of birth, 73% of 2011 census respondents were linked to their responses from the 2006 census. This included responses from the 2006 long-form for a large portion of the 2011 NHS sample, including non-respondents to the survey. This information was used to compare several characteristics of respondents and non-respondents to the NHS that tended to be stable over time, and to produce and analyze bias indicators in order to evaluate the quality of the estimates derived from the survey. The limitation of this approach is that not all records from the NHS could be matched this way. Also, the production of the bias indicators was only possible for larger geographic areas such as provinces and territories and Census Metropolitan Areas. Finally, the analysis excluded all new residents to Canada

who came after the 2006 Census. The bias indicators were exactly as the name implies, indicators of bias and not a measure of bias.

10. Results

Because of data quality, Statistics Canada is releasing less data from the NHS than with the mandatory long form of 2006 and previous censuses. All the 151 Census Metropolitan Areas and Census Agglomerations in Canada (these are the largest geographical entities below province and territories) meet the 50% GNR threshold for dissemination. Of the 4,567 Census sub-Divisions (CSD) with a population of 40 people or more, 3,439 (or 75%) met the set GNR threshold for release in standard data products. Only 245 CSDs with a population of 40 or more were not released because of data quality for the 2006 Census. CSDs represent municipalities or rural unorganized areas. Nothing is released for CSDs with a population less than 40 for confidentiality reasons. The CSDs meeting the quality threshold represent 96.6% of the Canadian population. The province most impacted by quality suppressions is Saskatchewan with 57.5% of CSDs released. Most CSDs not meeting the release threshold tend to be small and located in rural areas.

For the regions released, there is more volatility in estimates compared to previous census long-forms. This is due to the higher weights caused by higher non-response. Weights for the Census were rarely over 10, which rarely caused records with rare characteristics to stand out in overall results. Weights for the NHS can be as high as 100. In smaller areas, this causes some records to have a disproportionate impact on results on specific characteristics, rendering comparisons on trends difficult. Observations not based on a minimum of 4 unweighted observations are indicated as 0 in any tables to avoid direct or residual disclosure risks.

Finally, there are also results at the national level that are not in line with what might have been expected. This conclusion was reached in data certification when NHS estimates are compared to other sources, such as other Statistics Canada surveys that have shown consistent trends. Information on such occurrences are available on Statistics Canada's web site with the analytical and data outputs released for the survey.

11. Conclusion

The Canadian experience with the 2011 National Household Survey shows that a voluntary survey as designed cannot meet all the requirements for small area and small population data. Efforts described in this article went a long way to attenuate the potential for bias. The approach and methodology for a survey should always be set in function of the results that are expected. It is the case for all surveys at Statistics Canada, whether conducted on a mandatory or voluntary basis. The NHS, like the Census detailed form before, is the only sources of detailed standard data at the national level for small areas. To attain an acceptable level of data quality for small area, a high response rate must be achieved.

Statistics Canada is currently analyzing the lessons from the 2011 survey to plan for the 2016 Census program.

Reference

Statistics Canada, National Household Survey Users Guide, May 2013