Developments in internet collection at Statistics Canada

Claude Julien
Statistics Canada, Ottawa, Canada
e-mail: Claude.julien@statcan.gc.ca

Abstract

Statistics Canada has successfully used internet collection in its 2006 and 2011 censuses. Building on this success, it has started to roll out internet collection in its household and business surveys, including in its monthly Labour Force Survey (LFS). Concurrently, it has combined various sources of information to augment its Address Register into a Household Survey Frame Service (HSFS). This paper provides an overview of internet collection at Statistics Canada, some of its challenges, an experiment to assess the effectiveness of new collection strategies using internet collection, and recent initiatives that may further strengthen the HSFS to make more efficient and effective use of internet collection in the future.

Keywords: non-response error, frame, household surveys.

1. Introduction

The internet is increasingly being used to conduct surveys. For censuses or probability surveys, it can be used as a mode of data collection (internet collection) and is often combined with other modes (telephone or in-person). It can also be used as the sole source of information (internet surveys). In both cases, it offers a cost-efficient means of collecting data for certain segments of the population. For internet surveys where individuals volunteer to answer, even in very large numbers, biased estimates can be an outcome.

Internet collection, even on a probability sample of individuals, is prone to biases as well. First, response rates can be quite low, sometimes comparable to traditional mail-out surveys. For this reason, internet collection must be supplemented by other means of collection such as telephone or personal visit. Secondly, internet collection poses additional challenges when an individual must be selected at random among members of a household.

There are means of mitigating the risks of biases. In the context of internet collection, a subsample of non-respondents can be contacted by another mode of collection. In the context of internet surveys, Rivers (2007) proposes a technique to confer probabilistic characteristics to non-probability sample. Either internet collection or surveys can be partly corrected with poststratification or calibration techniques. A good quality frame, from which to select a sample or subsample, to plan collection, to match or to form calibration classes is essential to ensure efficient and effective use of internet technology.

National statistical offices have access to many sources of information and use them extensively to create and update frames. Some countries have formal population registers; other countries, such as Canada, have combined several sources of administrative data to create statistical registers or frames.
This paper first describes the current and future use of internet collection at Statistics Canada. It summarises the results of an experiment that was conducted to assess means of selecting random individuals when using internet collection in a household survey. It continues with recent and future developments to strengthen our household survey frame infrastructure to be in a better position to make more efficient use of internet collection and experiment with internet surveys.

2. Internet collection at Statistics Canada

Statistics Canada is increasingly using the internet as a collection medium through electronic questionnaires (EQ). It is the primary mode of collection in its Census of Population, National Household Survey (NHS) and annual business surveys carried out in the Integrated Business Statistical Program (IBSP). A few months ago, it was introduced in the monthly Labour Force Survey (LFS).

These surveys all use other modes to collect data for some parts of the target population, or to follow-up non-respondents. The NHS is voluntary and the other three are mandatory. As part of its Integrated Collection and Operation System (ICOS) initiative, Statistics Canada will be rolling out EQ collection in most of its household surveys. Many other annual and sub-annual business surveys will be transitioning to EQ as they are moved to the IBSP infrastructure.

The Census is mostly based on using an Address Register to contact over 80% of all dwellings with an invitation to fill the census questionnaire on-line; the rest are contacted in-person (Hamel and Béland 2013). The three surveys are based on probability samples selected from well established frames: Census of Population (NHS), business register (business surveys), and list of area units supplemented with an address register (LFS). This context is far from internet surveys. Yet, declining response rates in government surveys, especially among certain segments of the population bring about challenges that are similar to those faced by internet surveys. Self-selection of respondents, be it at the core of the survey design or introduced in the response process of a random sample of individuals may lead to biased estimates.

3. Selection of random sample of individuals in internet collection

The Census, NHS and LFS collect data from all or most members of a household. Many other household surveys are designed to select and contact a random sample of individuals. The current typical approach consists of selecting a random sample of dwellings through a list or area frame, and to contact the occupants of the selected dwellings by telephone or in-person. The interviewer starts by listing every occupant in the dwelling and their basic demographic characteristics. From this information, an individual occupant is selected by some random process.

To make the fullest use of internet collection requires a means of assuring the selection of a random sample of individuals without the assistance or intervention of an interviewer. In particular, we would want to avoid having the surveys completed by the most available and willing person in the household which would bring them closer to the limitations of internet surveys.

In early 2014, we conducted an experiment to measure the effectiveness of new strategies using internet collection for household surveys. In this test, respondents were only contacted through mailed invitations. Two strategies were tested to select individuals at random:

1. Household strategy: Mail letters to a sample of dwellings with an invitation to do the survey via internet collection (EQ). The invitation was addressed to “the occupant”. Once online, the household respondent was asked to complete a household roster, at which point one person 15+ in the household was selected to complete the questionnaire. If the selected respondent was the same person who completed the roster, the EQ application continued to the content; otherwise, the EQ application ended and a new invitation to the named selected respondent was sent.
2. Target Respondent strategy: Mail letters directly addressed to a sample of individuals, i.e. using their name, with an invitation to the EQ survey. With this strategy, there was no need for a roster. Two versions of invitation letters were tested under this strategy; one with the same letter used for the household strategy; the other letter had one additional

sentence stating that the respondent may be contacted by an interviewer if they do not complete the questionnaire.

Collection lasted approximately six weeks and took place in January and February 2014. Follow-up efforts were restricted to sending reminder letters. The survey content was on social identity and internet use. The surveyed population for both strategies was extracted from households that had completed the 2011 National Household Survey (NHS). The sample size for the Household strategy was 30,000 dwellings. The sample for Targeted respondent strategy was 6,500, which was split into two samples – 5,000 and 1,500, where the smaller sample received the alternate introductory letter.

The overall outcome indicates that it is not clear that one strategy is better overall than the other. The household strategy had the advantage of having many fewer post office returns (2%, due to unoccupied or inexistent dwellings) over the target respondent strategy (12%, due to selected persons having moved since the 2011 NHS conducted over 2.5 years earlier). Putting the post office returns aside, both strategies had 30% of the selected dwellings with an occupant log into the restricted access EQ application. In the case of the target respondent strategy, the occupant was the selected person and, thus, this 30% is the log-in rate. The alternate letter rendered a higher log-in rate (34%). For the household strategy, the log-in rate was significantly lower at only 22%. It was due to a loss in having to re-contact households when someone other than the initial responding occupant was selected and failed to log back onto the EQ application to complete the questionnaire.

Not surprisingly, neither strategy could be used as the sole mode of collection. Both strategies provide a certain percentage of respondents at low cost, but they must also be combined with non-response follow-up by telephone or in-person. Of the two strategies, targeting selected respondents would be more efficient if we had more accurate and up-to-date information on households. The next section describes recent Statistics Canada initiatives (and combination of distinct initiatives) that have produced a stronger household survey frame infrastructure and have the potential of further enhancing it in the years to come.

4. Strengthening the household survey frame infrastructure

Up until the mid-2000s, most household surveys at Statistics Canada were designed mainly on an area frame or a telephone frame (random digit dialling – RDD methodology). Over the years, the area frame has been increasingly supplemented by lists of dwellings extracted from Statistics Canada's Address Register (AR) and the RDD methodology has been abandoned.

The AR was first developed for the 1991 Census of Population and was used until the 2001 census as a post-drop-off coverage improvement tool. Since the 2006 Census, it has been used to create the list for the mail-out of questionnaires or invitations to fill questionnaires on-line in census mail-out areas. It is updated with various administrative files that indicate potential new addresses and presently covers 98% of all dwellings in Canada. To fulfill the needs of household surveys, it has been combined with numerous sources of telephone numbers and with socio-economic data into a Household Survey Frame Service (HSFS). As of the end of 2012, the HSFS comprised nearly 16 million dwellings, of which nearly 87% had at least one telephone number; over 96% were either reachable by telephone or mail; over 83% had information on the household size and composition (age and sex of members); and over 84% had some information on income. The addresses and telephone numbers are updated on a quarterly basis, and most of the socioeconomic characteristics are updated on an annual basis.

The socio-economic information comes from the 2011 census and tax data. The latter are provided by the T1 Family File (T1FF) program that processes individual tax filers (T1 declarations) and their dependants to disseminate small area socio-economic data on Canadians and their families (Statistics

Canada 2014). This program has been conducted annually since the early 1980s. The evolution of the tax program and the T1FF methodology over the years has brought it to cover over 33 million Canadians in 2011 which represented approximately 96% of Statistics Canada’s population estimate on Census day.

Statistics Canada has also launched two other important initiatives that, while not related to the HSFS, will likely produce outcomes that have the potential to further strengthen it. The first is to conduct research to examine the potential use of new and existing sources of administrative data to support our census program. This research builds on the knowledge of data and methods used to create the T1FF with other sources to improve the coverage (counts) and accuracy (location and characteristics) of the Canadian population with administrative data. The other initiative is to create an environment that facilitates the linkage of files from various socioeconomic domains, such as education, health, labour, income and justice, to increase Statistics Canada’s capacity to conduct cross-sectional and longitudinal analysis without imposing burden on respondents. Regardless of the extent and speed with which these initiatives will meet their respective objectives, they have the potential to provide a wealth of information that could improve the completeness, accuracy and timeliness of socioeconomic characteristics that could be assigned to the dwellings on the HSFS.

This additional information would increase our capacity to develop more efficient sample designs from the HSFS. It could also increase our ability to maximise our collection efforts by adapting our collection methods to the selected household or individuals. For example, selected dwellings with higher propensity of responding to internet collection could be sent reminder letters before switching to another mode of collection. Other dwellings would be followed up more quickly by other means. The additional frame information could also be combined with the profile of low-cost internet respondents to select a more efficient subsample of dwellings for follow-up. Finally, the additional characteristics on dwellings would increase our capacity to match internet surveys to make them more representative, if and when we use such surveys.

5. Conclusions

The Internet has quickly become a necessary mode of collection in surveys. It provides numerous opportunities and challenges to the production of official statistics. Used as the sole mode of collection, it can incur significant bias; strategically used with other modes, it can increase response rates for some segments of the target population or allow to direct more follow-up efforts to other segments.

Internet collection or internet surveys pose stimulating technical challenges. Many potential solutions rely on the existence of a good quality reference population database in terms of coverage of units, and the amount, accuracy and timeliness of information. This information can be used to determine the most cost-efficient mode of collection to use according to the characteristics of the selected units, or to determine when to quickly switch to another mode of collection. This information could be used to adjust survey weights for non-response. Finally, it could potentially be used to make non-probability surveys more statistically representative.

A few years ago, Statistics Canada made a strategic decision to move to internet collection as its primary mode of collection in most programs. It has been a success in the Census of Population and business surveys, and it has started being implemented in household surveys. To support household surveys, Statistics Canada has created a Household Survey Frame Service (HSFS) that covers 98% of all dwellings in Canada, most of them having contact information (mail or telephone) and some socioeconomic characteristics that are updated on a quarterly or annual basis. Developments in other key initiatives have the potential of providing more information to the HSFS that could be used to make more effective and efficient use of internet collection to produce official statistics and, maybe one day, internet surveys to new statistics filling current data gaps.

References

Hamel, M. & Béland, Y. (2013). “Future developments on the Canadian Census of Population”, 59th ISI World Statistics Congress, Hong Kong, http://www.statistics.gov.hk/wsc/IPS094-P3-S.pdf

Rivers, D. (2007). “Sample Matching for Web Surveys: Theory and Application”, Proceedings of the Joint Statistical Meeting, Salt Lake City, Utah, 2007.

Statistics Canada (2014) “Annual Income Estimates for Census Families and Individuals (T1 Family File) - Family Data - User's Guide, http://www23.statcan.gc.ca/imdb-bmdi/document/4105_D5_T1_V11-eng.pdf