CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection 'New Sources and New Technologies' Session 5

14-16 October 2019, Geneva, Switzerland 19 September 2019

Modernizing Data Collection in Canada

Lise Rivais (Statistics Canada, Canada)

lise.rivais@canada.ca

Abstract and Paper

As traditional methods to collect data from households are becoming less effective and more costly, new innovative approaches are emerging and must be considered.

Around the world, traditional primary data collection methods are becoming less effective and require more effort to achieve satisfactory results for household surveys. Technological and cultural changes have increased collection costs, as establishing contact with respondents and gaining their co-operation now require more effort. As a result, response rates for many surveys are trending downward. Finding new innovative ways of collecting the data necessary to create insights is very important for national statistical offices if they want to remain relevant.

Statistics Canada has recently increased its emphasis on researching and introducing such innovative collection methods for household surveys. As a result, response rates have stabilized and costs have been managed effectively over the last few years. The first part of the paper will describe the initiatives that successfully contributed to alleviating the downward trend in response rates. However, continued research is required on new data collection methods and techniques, as the downward trend in response rates could return, along with resulting cost increases to limit it.

As a result, Statistics Canada is researching more advanced approaches, which might change its primary data collection more dramatically by complementing or replacing traditional collection. The next steps are thought to lead towards completely new data collection techniques, such as sensor and scanner use, crowdsourcing, web scraping, automated voice interface use, and other innovative methods. The second part of the paper will describe some of the experiments, risks and opportunities that are being considered at Statistics Canada. It will also provide suggestions to identify and consider even more innovative and modern approaches.

Keywords

Costs; innovative; experiments; timeliness; relevance

Modernizing data collection in Canada

*Stéphane Dufour[1] stephane.dufour@canada.ca, Geoff Bowlby[1] geoff.bowlby@canada.ca, François Laflamme[1] francois.laflamme@canada.ca, Sylvie Bonhomme[1] sylvie.bonhomme@canada.ca and Holly Mullin[1] holly.mullin@canada.ca. Contributions from Etienne Saint-Pierre[1], Fred Barzyk[1] and Sevgui Erman[1].

[1] Statistics Canada, Ottawa, Canada

Keywords:
Costs; innovative; experiments; timeliness; relevance

As traditional methods to collect data from households are becoming less effective and more costly, new innovative approaches are emerging and must be considered.
Around the world, traditional primary data collection methods are becoming less effective and require more effort to achieve satisfactory results for household surveys. Technological and cultural changes have increased collection costs, as establishing contact with respondents and gaining their co-operation now require more effort. As a result, response rates for many surveys are trending downward. Finding new innovative ways of collecting the data necessary to create insights is very important for national statistical offices if they want to remain relevant.

Statistics Canada has recently increased its emphasis on researching and introducing such innovative collection methods for household surveys. As a result, response rates have stabilized and costs have been managed effectively over the last few years. The first part of the paper will describe the initiatives that successfully contributed to alleviating the downward trend in response rates. However, continued research is required on new data collection methods and techniques, as the downward trend in response rates could return, along with resulting cost increases to limit it.

As a result, Statistics Canada is researching more advanced approaches, which might change its primary data collection more dramatically by complementing or replacing traditional collection. The next steps are thought to lead towards completely new data collection techniques, such as sensor and scanner use, crowdsourcing, web scraping, automated voice interface use, and other innovative methods. The second part of the paper will describe some of the experiments, risks and opportunities that are being considered at Statistics Canada. It will also provide suggestions to identify and consider even more innovative and modern approaches.

1. Introduction:

Statistics Canada, like many statistical organizations throughout the world, has observed a downward trend in household survey response rates. Changes in the external environment (e.g., more cellphone-only households) and changes in respondent behaviour and their communication preferences have led to this steady decrease. Statistical organizations are asking: what types of initiatives can improve response rates?

Statistics Canada has responded in two different phases.

The first phase, almost completed, focused on better managing current collection approaches. This phase saw most effort devoted to developing an electronic questionnaire platform that enables web-based and multi-mode data collection strategies, answering respondents' demand for more convenient electronic self-reporting modes. In addition, Statistics Canada has made important moves towards improving the management of cases in collection (such as case prioritization and implementation of responsive collection design (Laflamme et al. (2016)), improving the allocation of interviewers' workloads and managing survey operations more actively.

Statistics Canada recently began the second phase of its research, focusing on new data collection methods and techniques that might be more aligned with respondent preferences and reduce respondent burden. These new primary data collection modes aim to be easier to use, more efficient and less burdensome than the usual collection approaches, or even eliminate the need for surveying altogether. This paper also seeks to provide, in Section 2, an overview of Statistics Canada's recent successes in better managing its alternative data collection process and practices. It also briefly describes Statistics Canada's new data collection initiatives.

2. Phase 1: Better management of current collection approaches:

This section presents an overview of initiatives that have been successfully implemented and that have contributed to alleviating the downward trend in response rates.

New e-questionnaire platform

Survey respondents in Canada increasingly expect an electronic self-response mode. Some years ago, Statistics Canada set out to build this option for its respondents, while at the same time replacing a myriad of data collection systems that were becoming increasingly difficult and costly to maintain. The resulting Integrated Collection and Operation System was first used for the 2016 Census of Population and later adapted for use by all business, household and agriculture surveys, as well as for Consumer Price Index data collection.

This new system has resulted in approximately 80% of Statistics Canada's surveys now offering an HTML-based, multi-mode-ready questionnaire, which can be delivered to a respondent's computer, laptop or other mobile device, and which can also be accessed by interviewers in homes or in a call centre. The remaining 20% of surveys are planned for migration to the new system within the next 24 months.

The new e-questionnaire platform is achieving two goals. The first is to provide respondents with their preferred response mode. The second is cost savings, since the self-response mode is reducing the hours of interviewing required. The estimated annual savings from offering an e-questionnaire option (not including the census) are CAN 2.9 million so far.

Case prioritization and interviewer allocation

Case prioritization was developed to improve sample representativeness by targeting high-priority surveys or cases that belong to domains with lower response rates. In some circumstances, case prioritization might be used to target specific cases for various operational reasons. The objective is to monitor data collection while it is in progress to identify the cases to prioritize. It is one of two "adaptive" approaches (the other being responsive collection design), which use information available before and during collection to adjust the collection strategy for the remaining in-progress cases.

As part of the recently deployed collection platform, rules to deliver cases according to the highest priority have been introduced to govern work in Statistics Canada's five call centres. These rules have various levels. For example, at one level, the rule assigns cases to call centre agents so that they work only on a given survey, or in proportions x, y and z on several given surveys. The call centre would pay attention to these particular surveys on that day. Next, the prioritization system targets specific operations, such as non-response follow-up or refusal conversion, within the priority surveys.

The allocation of interviewer efforts is related to case prioritization. Research at Statistics Canada had shown that staffing levels were not always well aligned with the workload sample and expected productivity (Laflamme (2008a); Laflamme (2008b)). In response, Statistics Canada has optimized interviewer efforts on cases where they will be more efficient.

Another initiative is to automate the delivery of specific cases, such as those eligible for responsive collection design.

Responsive collection design

Responsive collection design (RCD) is a technique Statistics Canada has used in production for all computer-assisted telephone surveys since January 2015, following a series of experiments in previous years.

Using RCD at Statistics Canada resulted in higher response rates and improved data quality, without increased costs or burden to Canadians. A typical RCD approach divides the collection operation into phases. The earliest phase begins the survey with a traditional, randomized allocation of questionnaires to interviewers. Next, the interviewers are asked to complete certain cases that are more likely to bring about a successful response. The final phase emphasizes more difficult cases to reduce the differences in response rates between the domains of interest.

Active management

Traditionally, Statistics Canada surveys have been managed through the regional offices, where all survey taking takes place. Operations management was left entirely to those offices, with relatively little planning and support from the central office located in Ottawa. Recently, Statistics Canada changed this approach and introduced a new set of common plans and tools to centrally manage survey data collection in progress through an active management unit.

The active management unit in Ottawa has three main objectives. The first is to determine data collection milestones where changes to the collection strategy are required. The second objective is to identify problems as early as possible and correct them (if required) before collection has finished. The third, which is a more global objective, is to use collection resources effectively to find the most appropriate balance between data quality, timeliness and survey costs. Active management is based on current, timely and empirical observations, and it is considered one of the main reasons Statistics Canada's response rates have stabilized.

Active management is best demonstrated by explaining the tools that managers now have at their disposal as a result of this program. For any given survey, all data collection managers have access to a national production plan before collection (which they have an opportunity to influence). They also have access to monitoring reports delivered centrally on a regular basis. The monitoring reports come with basic analytical information designed to identify collection issues and potential solutions to any issues that are noticed through monitoring.

The plan for active management is to refine available tools, including by implementing data visualization tools, and to continue to work towards establishing operational survey "command centres" in each region and at headquarters in Ottawa. The goal is to improve responsiveness and optimize data collection resources.

Expansion and improvement of respondent communication material

About five years ago, Statistics Canada focused on communication material to improve response rates. The idea was to "nudge" respondents using the latest research on behavioural economics and show, through various tools and integrated activities, that survey participation is useful. A framework was developed to prioritize needs for communication support based on the type of survey, the importance of the survey and the expected response rate.

The best example of an effective new communication strategy is the one used for the 2016 Census of Population. Based on results from the previous Census of Population, the population of Canada could be divided into five different groups, each with its own unique communication strategy. The groupings were based on the likelihood of a fast response to the census. One group, the easiest to reach, got relatively light communication. People who are more difficult to reach, on the other hand, received communications at various stages of collection. This segmentation strategy is an important reason why the 2016 Census of Population was considered the best ever in Canada, with the highest response rate on record, and with an impressive cost and quality performance.

3. Phase 2: Experimenting with new data collection methods

As mentioned earlier, focusing only on optimizing Statistics Canada's current collection operations would be insufficient. New ways of collecting data must be explored to reflect the new reality of a population less interested in completing surveys and to take advantage of technologies now available that could transform primary data collection operations. This section presents Statistics Canada's research focus areas, considering the anticipated operational implementation feasibility.

Developing a crowdsourcing service

Crowdsourcing has been an early success in the introduction of new primary data collection techniques. Crowdsourcing involves asking the population to proactively provide information rather than wait to be contacted when selected as a respondent. The risk of such an operation is obvious to the statistician—crowdsourcing data quality is difficult to assess, with metrics on quality near-impossible.

Nevertheless, Statistics Canada began to experiment with crowdsourcing in areas deemed relatively low-risk. The technique was first used for a project to improve the available information about dwellings in Canada. The "crowd" was asked to provide GPS locations for a set number of dwellings posted on the Statistics Canada website. That drew considerable interest from the population, who provided the requested information faster than expected. Secondly, Statistics Canada crowdsourced the price of cannabis through its StatsCannabis web application in the months before the legalization of cannabis in Canada in fall 2018, when cannabis consumption was still illegal (except for approved medical use). This resulted in over 20,000 entries to the questionnaire, and reasonable price estimates (i.e., within expectations, upon validation).

Subsequently, Statistics Canada implemented a crowdsourcing service within its survey operations branch. The service is relatively simple—a short e-questionnaire, accessible to all website visitors (i.e., there are no barriers such as access codes to keep people from responding). For each crowdsourcing operation, there is a structured communication plan aimed at specific crowds. Active monitoring processes are also used to ensure projects are successful. In the last eight months, a number of new crowdsourcing operations have begun in Canada, mostly to collect qualitative information for Statistics Canada, as it designs new products and services. Public participation in Statistics Canada crowdsourcing has shown successful results.

Using SMS messages as a reminder to respondents

Statistics Canada is currently piloting the use of SMS (short message service) as a survey reminder strategy in an effort to encourage respondents to report their data after their initial invitation. This is part of a strategic plan to take a more user-centric approach to contacting respondents.

The first pilot survey saw 13,000 respondents receive text messages as their fifth (last) reminder, and it generated a response rate of 1.5%. This compares with the usual take-up rates of 1% for paper (mail) reminders and less than 1% for email follow-ups. More research is planned to assess the impact of using SMS for earlier reminders (first, second or third) to improve comparability with other modes and evaluate the cost effectiveness of this new way to contact Canadians.

The pilot survey was implemented in partnership with a major telecommunications company in Canada, using an SMS aggregation system. This tool enabled Statistics Canada to send automatic mass communications effectively at a reasonable cost of CAN 0.50 cent per SMS.

Before testing the use of SMS, Statistics Canada engaged in extensive public consultation. This consultation revealed that respondents are likely to view an unsolicited SMS from Statistics Canada (i.e., cold call via text) negatively, but would find it acceptable if they had already been in contact with Statistics Canada through another mode and were informed about the potential use of SMS. In addition, Statistics Canada consulted with the Office of the Privacy Commissioner, which reiterated the recommendation that Canadians be given advance notice that they might receive an SMS. This is why Statistics Canada has chosen to use SMS for reminder notices, rather than earlier contact with respondents.

Using a Statistics Canada data collection application on mobile devices

Statistics Canada is currently investigating the use of a mobile application to collect data for household surveys that require respondents to report information several times a day or on several days. This would give respondents a readily available, user-friendly collection tool for surveys that require repeated input, such as time use surveys or consumer expenditure reporting.

In addition to providing a convenient way for respondents to complete some of the most burdensome surveys, an application could take advantage of the option to ping respondents at strategic points in time to nudge them to respond. It could also benefit from the multiple sensors that smartphones currently use, including GPS and pedometers, as well as any connected devices with sensors, such as Fitbits. Data collected elsewhere on the device could be used by the application if the respondent permitted it and if the data suited the project using the app.

A first pilot project is planned for 2019/2020: a new survey measuring subjective well-being in Canada. Because of potential privacy issues, Statistics Canada is currently studying the legal and IT risks before completing specifications for the required solution and is working in collaboration with the Office of the Privacy Commissioner and IT security experts.

Testing cognitive interactive voice response technology

Later in the 2019/2010 fiscal year, Statistics Canada anticipates testing cognitive interactive voice response (IVR) technology as an interviewer or respondent monitoring tool. Cognitive IVR is a way for humans to interact with an artificial intelligence platform, such as IBM's Watson, Google Duplex and other similar products. This can be done with a number of methods, but as a first step, Statistics Canada is planning to explore ways to automate quality control and interviewer feedback through a cognitive IVR system. In addition, Statistics Canada hopes that the instant feedback of these platforms will enable it to successfully collect more data by being able to better tailor its tone and approach to each individual respondent.

This would be a first step towards using cognitive IVR in a way that would have bigger consequences for Statistics Canada's data collection operations. In addition to providing automated feedback to human phone operators, this technology could enable respondents to call a phone number and be interviewed by the cognitive IVR system, just like they would with an interviewer. This could be used for all or part of an interview process.

Using scanner data to collect information

Statistics Canada has completed the first year of a three-year plan to introduce scanner data into the production of the Consumer Price Index (CPI). The ultimate goal is to replace all traditional food price collection in the field (in-store collection). A simple implementation plan has been put in place whereby each field-collected quote is being replaced by an average price for the same or similar product using scanner sales data. Statistics Canada has successfully integrated one major retailer and is scheduled to introduce two more in the fall of 2019. Savings from the reduction of field collection costs are being reinvested into the development of tools necessary to support collection from alternative data sources such as scanners.

This new source of data required new processing tools to be created outside the traditional system. The scanner data need to be pre-processed to link the products to the CPI classification—machine learning is now being used for this classification process.

Further developments are being considered, such as automated substitution of products and the use of multilateral methods of index calculation (i.e., the use of all data over time rather than a sample that mimics field collection). Both of these developments are beyond the scope of the three-year plan and would not be implemented before 2021.

Using sensors to collect information—satellite and telemetry

Satellite imagery is a key component of the Agriculture Statistics Program. Statistics Canada has collected vegetation index values from satellite imagery since the 1990s to support the Crop Condition Assessment Program, a web mapping application that depicts crop and pasture conditions across Canada in near real time. This data source, coupled with climatic data, is the foundation for the crop modelling project. The results of this project have been accurate enough to replace traditional collection methods for the September Field Crop Survey since 2016, eliminating more than 9,000 phone interviews. In 2019, Statistics Canada is looking to expand the modelling approach of the Field Crop Survey to further reduce response burden.

In addition, Statistics Canada successfully used a combination of crop insurance data and a crop classification map produced by Agriculture and Agri-Food Canada with medium-resolution satellite imagery to estimate crop area. Results were primarily used for validation at first, and this method is another potential way to reduce response burden for the Field Crop Survey and future censuses of agriculture. A 2019 pilot project is looking into updating crop area and yield on a weekly basis, at the parcel level, as the growth season progresses (in-season estimates) using near real-time satellite imagery, climatic data and crop insurance data.

Statistics Canada is also working on an innovative project with the Canadian Food Inspection Agency using the traceability data managed by the Canadian Pork Council, which collects group movement data. Statistics Canada is using data science and leading-edge methods to clean, process and use the data to create real-time modelled pig inventories by location, along with the probabilistic movements of each animal in the group. This partnership on traceability in the pork industry is an excellent opportunity to benefit each organization's goals. Statistics Canada will have information on pig movements and the inventory required for its statistical programs. The Canadian Food Inspection Agency will benefit by having a framework on disease predictability to help monitor potential outbreaks of disease, such as African swine fever.

Exploring web scraping as a new mode of collection

Currently, Statistics Canada's Annual Survey of Manufacturing and Logging Industries uses available information to prefill components of the annual questionnaires to facilitate reporting of the commodities being produced, as well as the sales amount. The commodity data are prone to non-response and reporting errors but are crucial for measuring economic production in Canada. A pilot project will investigate the use of web scraping technology to collect website information to get a better picture of the types of commodities manufactured and to potentially improve sales estimates. A generic web scraper will be used to collect text data, which will be transformed into a list of commodities from each company's website. This information will then be used to prefill the questionnaires described above and improve the high non-response. In addition, the data will be used to improve auto-coding rates and quality.

Automated web scraping involves important ethical and privacy issues for national statistical offices. In response, Statistics Canada is developing a directive on web scraping to establish recommendations about using API when available, consulting the robots.txt of websites, respecting website controls and protocols, not capturing personal information, being transparent, and addressing other issues. All of this work will be done in close collaboration with the Office of the Privacy Commissioner.

Statistics Canada is also exploring web scraping from news platforms to collect information that can support enterprise profiling, financial variable coherence analysis, merger and acquisition event detection, and sentiment indicators for measuring business tendencies based on news analytics.

4. Conclusion:

While it may be a challenge for national statistical offices to collect data with response rates declining around the world, Statistics Canada has implemented changes to stabilize response rates and adapt to respondents' preferences in terms of collection or contact mode. In addition, Statistics Canada feels that the new tools and techniques that are now available show great promise in supporting its work. These changes and techniques were summarized in this paper. Statistics Canada is happy to partner with other national statistical offices interested in experimenting with and developing such solutions, to improve the future of data acquisition.