Evaluating Data Quality in a Large Voluntary Survey: Canada's National Household Survey

Sander Post

Statistics Canada sander.post@statcan.gc.ca

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Introduction

Canada conducts a census of population and housing every 5 years. The Census is a based on a dwelling frame – we create a list of all dwellings, and then enumerate the usual residents of each dwelling. Historically, eighty percent households (except for those living in remote areas and Indian Reserves) were given a short form that asked for basic demographic information of each resident. The remaining twenty percent were given a long form that, in addition to the short form questions, also asked more detailed questions about topics like occupation, education, ethnicity, as well as about the dwelling. Both the short form and the long form were mandatory.

For the 2011 Census, the short form questions remained mandatory, and a short form questionnaire was sent to all households. Long form questions were asked as part of a separate survey, the National Household Survey, or NHS. This survey was voluntary, and the sample fraction was increased to one in three.

The decision to make the survey voluntary was made in 2010. The change was taken after Statistics Canada conducted field tests in 2009 based on the planning assumption that the long form would be a mandatory survey. Consequently, no data existed on which to estimate what impact this decision would have on response rates. In particular, we could not estimate what NHS response rates would be, and, consequently, what impact this would have on the data quality of this survey.

As could be expected, the response rate to the voluntary survey was lower than the response rate of previous mandatory surveys with similar content. In 2006, the response rate to long form questions was around 94%. By comparison, the short form had response rates of 98% in 2006. In 2011, the weighted response rate achieved for the NHS, after using a subsampling approach to sample and target initial non-respondents, was 77%.

To account for nonresponse, the design weights were adjusted using responses to the ten 2011 Census questions (i.e. age, sex, families, households and marital status, structural type of dwelling and collective dwellings, languages) and geographic information. To enrich the information used for the non-response adjustment, several probabilistic linkages of administrative files were done to both the census and the NHS. Linkages were done for the 2010 income tax data, the 2011 Indian register and the immigration database. For more information on weighting see Verret (2013).

Voluntary surveys have known quality concerns. In particular, if nonresponse is correlated with any variables, those variables may be under or overestimated in the final results. Consequently, we wanted to measure the impact of nonresponse on our estimates, and to see to what extent existing data would allow us to measure nonresponse bias for each variable. As part of that, we wanted to know if there were levels of nonresponse above which we should suppress data due to lower quality.

Part of our job as a national statistical agency is to provide guidance to users of the data. As users, particularly provinces and municipalities, were accustomed to data from a mandatory survey, we wanted to be able to make quantitative statements about the quality of the data we released, and provide information about why we suppressed data if we needed to.

This paper discusses approaches used to measure the impact of nonresponse, and how we used those results.

Record Linkage

One approach to measuring the impact of nonresponse is to obtain the missing data for nonrespondents from other sources. If this data is available, we can measure if nonrespondents differ from respondents, and thus measure to what extent our published estimates are biased. One readily available source of data is the 2006 Census. If we can link an NHS nonrespondent to a 2006 Census response, in particular a long form response, we can obtain their data. The goal would be to link to the 2006 Census basic demographic data (i.e. short form data), and then use long form data for the fraction that received the long form.

There are multiple uses for such a record linkage, and we were fortunate that another project at Statistics Canada saw fit to do this work for their own requirements. In particular, Statistics Canada conducts a number of post-Censal (and now, post NHS) surveys where target populations are found among long form respondents. To see what the impact was of nonresponse on the universe for these post-NHS surveys, they needed to evaluate the impact of nonresponse.

Under many circumstances, it would be difficult or impossible to link non-respondents to other data. Basic information, like the names of occupants, or their age and sex are, by definition, unknown. For the NHS, this is not actually the case. As the decision to go with a voluntary survey was taken at a relatively late stage in the survey process, we continued to use the already developed survey infrastructure. One result is that the Census and the NHS used the same frame at the same time, ensuring that household membership was the same. Consequently, for NHS sampled units, whether they responded or not, we would have had Census data for them in virtually all cases. Consequently, for NHS nonrespondents, we actually knew the number of occupants of the household, and the name, age, and sex of each occupant via a link to the Census.

The next step was to link all 2011 Census respondents to 2006 Census respondents, and obtain both 2006 long form and 2006 short form data.

Covering all the details of record linkage is beyond the scope of this paper, but we will list some obvious issues. Grenier (2012) contains details about this linkage and the issues encountered. The variables used at different steps in the linkage process include date of birth, sex, name, and phone number. Geographic variables were used to restrict the set of data to match to. For NHS respondents, the mobility question further helped this process, as it told us where the respondent lived 5 years ago, in 2006.

Not all records can be linked. Of the 2011 population of Census respondents, only some are in scope for linkage to 2006. Those not yet born in 2006, and immigrants who came to Canada after 2006 are both out of scope. Similarly, we cannot link to people who were nonrespondents in 2006, as there is no data to link to. One feature of the NHS did make linkage easier for NHS respondents; one of the questions is about mobility, namely, where did the respondent live 1 and 5 years ago. Thus, for respondents, we know where in the country to look for their response in the 2006 Census. As our goal was, for each 2011 respondent, to link to a unique 2006 respondent, other cases become difficult to link uniquely – some names are common, and some people were counted at multiple households in 2006 and/or in 2011. All of these cases make linkage difficult. For purposes of this study, if there was ambiguity in a linkage, we did not make the link.

About 73% of 2011 Census respondents were linked to 2006 Census respondents. To quantify the upper limit of matching, the population of Canada as of 2011 was 33.5 million, and as 3.1 million would have been out of scope (births and immigrants), our upper limit of matching would have been 90.7%. Consequently, about 80% of matchable respondents were actually matched.

How did we get permission to link respondent data or non-respondent data? The short answer is that such permission is largely implicit. When Statistics Canada collects data for the Census and the NHS, there are several statements in the introductory paragraph of the questionnaire stating the legal uses to which Statistics Canada can put the data. In effect, we promise the respondent that we will only release data in aggregate and that we will protect individual respondent's confidentiality. We state that we will use data collected for two purposes, disseminating data for the survey in question, and to improve our internal processes for future surveys. Research into analysing the results of a new survey certainly falls into the category of improving our internal processes.

Although the questionnaire does state that we can use data to improve our processes, researchers at Statistics Canada who wish to link records must follow a process to get approval to do so. A committee reviews proposals for studies requiring record linkage, and these proposals must include the planned uses for the results. Such plans must describe if the intent is to produce publishable statistics, or to measure some aspect of the survey, such as mode effects in a survey where a new mode of collection was recently introduced. The committee then approves or rejects the proposal, and if approved, may place restrictions on what can be done with the results.

Analysis

With the linked file, we can begin analysis. The NHS had approximately 55 questions in addition to Census questions. Before doing analysis, we wanted to select a set of questions we felt would have either stable answers, or answers that, while they could change, would change in predictable ways, or would infrequently change for certain demographics. As an example of the latter, respondent answers to the educational question of "highest certificate, degree or diploma obtained" can certainly change over time, but we believed that by using a subset – for example, people who were 30 or older in 2006 - we would have a fairly stable subset.

We started with the following set of variables:

Place of Birth
Place of Birth – Mother
Place of Birth – Father
Citizenship
Immigration Status
Year of Immigration
Visible Minority
Aboriginal Status
Registered Indian
Highest Certificate, Degree or Diploma

Before we used a linked file to measure nonresponse, we need to measure the quality of the file. In order, our main concerns are:

1. For 2006 linked records and 2006 unlinked records, are their responses similar?
    - We would be concerned if they weren't, as it would indicate a potential bias due to our linkage

2. Of linked NHS respondents, are answers similar in 2006 and 2011?
    - If answers are not similar, we should not use those questions to estimate nonresponse bias as there are clearly other sources of bias

3. Using information about 2011 NHS nonrespondents, can we generate indicators for nonresponse bias and can we use them to suppress data of poor quality?
    - This is the goal of the project

Regarding the first concern: when we examined 2006 linked and 2006 unlinked records, the only difference noted is that Aboriginal Peoples are linked at a higher rate. As part of our sample design is that Indian reserves (and some other remote communities) are sampled at 100% instead of being sampled at 1/3 or 1/5th, and they are collected by an enumerator instead of self response, consequently the observed differences in linkage rates are neither surprising nor of concern.

Before we calculate a bias indicator, we would like confirmation that responses are consistent over time. It would be difficult to calculate non-response bias indicators in the face of large amounts of other error – such as capture error, recall error, or errors due to proxy responses.

When we compared NHS respondents to 2006 respondents for some variables, we got the following results.

Variable	Matching Percentage	Unmatched Percentage
Place of Birth	98.6%	1.4%
Place of Birth – Mother	96.5%	3.5%
Place of Birth – Father	91.9%	8.1%
Citizenship	93.5%	6.5%
Immigration Status	98.5%	1.5%
Visible Minority	95.7%	4.3%

The primary result – these variables are over 90% consistent. This indicates that our matching is solid, and that respondents do tend to answer the same question the same way over time. We can, and do, speculate as to why some numbers are different, but have not researched all differences in depth. Place of birth should, in general, not change. It is possible some records were linked in error, and proxy response error and data capture error can likely explain the remainder. Still, a difference of 1.4% indicates that we can use this to calculate nonresponse bias indicators. Place of birth of mother or father have higher differences. We suspect this is partly due to mixed families – if parents divorce and remarry, do they list the biological parent or step-parent's place of birth? There are more proxy response issues – people would know more about the people in their household than about those individual's parents. Citizenship can change, and we examined this one more closely. For the most part, it went from non-Canadian citizenship to Canadian, which matches our expectations. One variable we assumed would be constant that wasn't particularly constant was year of immigration. Of course, years past, particularly distant years past, do pose recall error. But even when considered years to be a match if they were within 5 years, we still only got about a 60% match.

It should be noted that the wording of this question, year of immigration, was identical. However, in speaking with people who test Census questions, this one has often posed a problem. There has not yet been a wording developed that will not be misinterpreted by a surprising number of respondents. We looked more closely at this, and there are some cases where it is likely our automated data capture systems for paper questionnaires are causing errors, such as reading 1986 as 1936, but this is not widespread.

Nonetheless, with this analysis, aside from year of immigration, we believed that the selected variables were useful for calculating nonresponse bias indicators. They were either quite stable, or they would change in fairly predictable ways – in particular, citizenship and education.

To measure the impact of nonresponse, we need also to consider that responses are not static over time for all variables. We would expect that a person's place of birth should not change, nor would their date of birth. Other variables, such as educational attainment, can change, but should change in only one direction. And lastly, we have variables such as industry and occupation, which are neither static nor have any kind of predictable direction. This latter category is not of interest for this analysis.

All variables for matched respondents do show some change. However, we believe that for stable variables, the change is understood, and for variables that can change, the change is predictable. Based on this, we wanted to find a way to calculate the impact of nonresponse. Our approach was to develop a formula that adjusts our respondents and our non-respondents to a common total, and subtracts them, adjusting for the propensity of responses to change and for our weight calibration. Nambeu *et al* (2013) developed the following formula, which takes into account the propensity to respond, the likelihood of a variable to change, and the calibration done to the weights.

$$\hat{B}_2 = \hat{\beta}_0 \left(\sum_{i \in s_R} (w_i - d_i) f_{Ri} y_i^{2006} - \sum_{i \in s_{\bar{R}}} d_i f_{\bar{R}i} y_i^{2006} \right)$$

A full explanation of the derivation of this formula is available in the paper, but to describe the variables in order:

$\hat{B}_2$ is the calculated value of our indicator of nonresponse bias for a given estimate. For place of birth, for example, we calculated values for 100 categories.

$\hat{\beta}_0$ is an estimated parameter that indicates, for matched respondents, how the value of the variable changed between 2006 and 2011.

The first summation is over S_R, the set of matched respondents. Within the summation, we take the difference in calibrated weights (w_i) from design weights (d_i), multiplied by f_{Ri}, which is a weight adjustment factor accounting for the linked respondents, and finally by an indicator variable for 2006 – which was 1 if the record had the property (i.e. a given place of birth), and 0 if it did not.

The second summation, over the set of linked nonrespondents, included the design weight, the weight adjustment factor for the linked non-respondents, and the 2006 indicator variable.

As noted, in effect, the formula calibrates NHS respondents and NHS nonrespondents to the same total, and subtracts the values. Ideally, the values should be near zero. This would indicate that, after nonresponse adjustment, our estimates are the same as if our linked nonrespondents had all answered.

We calculated these at geographic levels from national down to municipal, results we used later in this project.

The following table shows the values we calculated for the top five places of birth in from the NHS.

Place of Birth	Bias Indicator	Total	Relative Bias Indicator
Canada	49304	18477127	0.3%
China, SARs, and Taiwan	1884	494556	0.4%
United Kingdom & Republic of Ireland	-4767	462939	-1.0%
India	6047	316498	1.9%
Philippines	19274	236486	8.1%

For these categories, the Bias Indicator shows how much we over estimated (positive numbers) or under estimated (negative numbers) due to nonresponse. The Philippines shows up as a value where bias is certainly evident, as our relative bias is 8.1%. This was in a note published with our data release that included the place-of-birth variable. As has been noted, this analysis is based on a linkage to 2006 respondents, and thus excludes post-2006 immigrants. However, additional comparison to data provided by Customs and Immigration Canada shows that for intercensal immigrants, our totals for immigrants from the Philippines also do not match the total. There has been reasonable sounding speculation as to why we are overestimating this category, but no definitive answer was known at the time of writing this paper.

While we do analyse data quality for its own sake, we had a specific goal in mind with the above analysis. Statistics Canada has historically suppressed Census data for two reasons: to eliminate the risk of violating a respondent's confidentiality, or to avoid disseminating data of poor quality. In previous Censuses, we have suppressed data for a very small number of municipalities where the response rate was considered to be too low. Given that the NHS is not the Census, and given that the new survey was voluntary, it was clear we needed to review our suppression rules. Using the indicators of nonresponse bias, we would like to measure the relationship between nonresponse bias and nonresponse rates, and, if suitable, use it for suppression of low quality data. Given data for approximately 5000 municipalities of various sizes, as well as higher level geographies, we could see if there was a positive correlation between nonresponse rates and nonresponse bias. If there was, we would choose a level of nonresponse as a cutoff to keep nonresponse bias within acceptable limits.

There was in fact such a correlation, and we used it to suppress data where the data was subject to high nonresponse bias. As it happened, at a municipality level, we released data for the vast majority of the Canadian population, but we also suppressed data for a significant portion of Canadian municipalities. The suppressed municipalities where generally of very low population, and were more often rural. We will use these results to improve and update our approaches to survey operations if we are to repeat a voluntary NHS in 2016.

Conclusions

The survey environments of 2006 and 2011 facilitated an analysis of nonresponse bias of the sort that cannot always be done in surveys. The results presented are, of course, only indicators. Nonetheless, for the in-scope population of this study, they allowed us to confirm that our weighting adjustments for nonresponse allowed us to produce quality estimates for variables we can reasonably study with this approach.

As there were nonrespondents in the group who are out of scope for linkage, in particular children born after 2006 and immigrants who arrived after 2006, we cannot, and have not, stated anything about the nonresponse bias in that group. However, comparison of NHS results to those from other administrative sources can show where our results seem to be biased for this group.

For the NHS in 2011, we had a mandatory base in 2006 to link to. In 2016, if a voluntary NHS is repeated, we will not have a mandatory 2011 base to link to, and consequently, we will not be able to do the same analysis as described in this paper. We will, however, have a track record of producing reliable estimates from the NHS, and we can build on that.

Acknowledgements

This research could not have been done without the work of many of my colleagues. In particular, I would like to thank Dominic Grenier and Chantal Grondin for creating the linkage file, Christian Olivier Nambeu for developing the formula that allows us to calculate indicators of nonresponse bias, as well as Laurent Roy and Scott McLeish for many contributions.

In addition, I would like to thank Jean-Pierre Morin and Normand Laniel for many invaluable comments that significantly improved this paper.

There are many others who contributed to this work.

References

Dominic Grenier, 2012. *Méthodes pour le micro-appariement des Recensements 2006-2011.* Internal Statistics Canada Document

Christian O. Nambeu, Normand Laniel and Sander Post (2013), *Indicateurs de biais de non-réponse de l'Enquête Nationale auprès des ménages.* Internal Statistics Canada document.

Statistics Canada (2013) National Household Survey User Guide, May 2013. Catalogue no. 99-001-X

Verret,F. (2013), "The estimation methodology of the 2011 National Household Survey". In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. p. 1876-1890.