The Estimation Methodology of the 2011 National Household Survey

François Verret
Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

Abstract

Prior to 2011, the Canadian census of population was conducted with a mandatory long-form sent to 20% of the households and a mandatory short-form sent to the rest of the households. The 2011 Census was conducted with a mandatory short-form sent to the entire population. A voluntary survey called the National Household Survey (NHS) was created to collect the information that used to be collected with the long-form. To minimize the impact of non-response, a sample of 30% was selected for the NHS and after several weeks of collection, the follow-up efforts were concentrated on a random subsample of the remaining non-respondents. The design-weighted response rate was 77%, while for the last census long-form the response rate was 94%. This paper describes the estimation methodology used in the 2011 NHS. Design weights were first calculated. A non-response weight adjustment was done using census and linked administrative data and by converting a nearest-neighbour and whole-household imputation approach to a reweighting approach. Weight calibration to many census totals was performed. Variance was estimated with a multi-phase variance formula and Taylor linearization.

Key Words: Large-scale surveys, subsampling of non-respondents, non-response adjustment, calibration

1. Introduction

The Canadian census of population has been conducted every five years since 1981. Before 2011, one in five households received a mandatory long-form while the remaining households received a mandatory short-form. In 2011, a mandatory short-form was sent to every household; however, the census long-form was replaced by a voluntary survey called the National Household Survey (NHS). The NHS is the largest voluntary household survey ever conducted by Statistics Canada. Although they were two distinct surveys, the 2011 Census and NHS were collected at the same time and using the same infrastructure. They also share the same reference date: May 10th 2011. Their target population differ nonetheless: the NHS covers households of all private dwellings, while the Census covers households of all private and collective dwellings.

The 2011 Census short-form consisted of ten questions. The topics covered were

- Population and dwelling counts
- Age and sex
- Families, households and marital status
- Structural type of dwelling and collective dwellings
- Language.

The NHS form consisted of the 10 census questions, 44 additional questions about the individuals and 10 questions about the dwelling. The topics covered by the 2011 NHS were

- The topics covered by the census (except for collective dwellings)
- Activity limitations
- Ethnic origin, visible minorities and immigration
- Aboriginal peoples
- Mobility and migration
- Education
- Labour
- Place of work and commuting to work
- Income and earnings
- Housing and shelter costs.

The risks associated with non-response are more important with a voluntary survey such as the NHS than they are with mandatory surveys such as the long-form censuses. Consequently, the sampling, collection and estimation methodologies that were used with the mandatory long-form were changed to minimize the effect of non-response in the NHS. This paper describes the estimation methodologies adopted for the 2011 NHS. In section 2, the NHS sampling design and corresponding design weights are presented. In section 3, adjustments due to the voluntary nature of the NHS are described. In section 4, the NHS weight calibration process is outlined. In section 5, the variance estimation methodology is given. The paper ends with a conclusion in section 6.

2. NHS Sampling Design and Design Weights

2.1 First Phase of Sampling

At the first phase of sampling of the NHS, the dwellings were stratified by Collection Units (CU). These small geographic entities were used to manage the collection process and contain on average 300 dwellings. To each CU corresponds a questionnaire delivery method. These entities respect provincial and territorial boundaries as well as Census Division (CD) boundaries (a sub-provincial and sub-territorial geography). In each of these strata, a systematic random sample of dwellings was selected. Sampling fraction varied by province and territory and by questionnaire delivery method. The overall sampling fraction was 30% and the sample size was approximately 4.5 million dwellings. Note that at the time of sampling it was not known if a given dwelling on the sampling frame housed a household that was in-scope for the NHS or not.

The first step of the estimation process was assigning a weight to each dwelling selected in the first phase NHS sample. The first phase weight assigned to a selected dwelling was equal to the inverse of the probability of selection of the dwelling and was the inverse of the sampling fraction in the stratum of the dwelling.

2.2 Second Phase of Sampling

After several weeks of collection, a random subsample of the remaining non-responding dwellings was selected and the follow-up efforts were restricted to this group of non-respondents. This subsampling approach was developed by Hansen and Hurwitz (1946) and can be pictured as in Figure 1. In this figure, the sampled population is represented by U, the first phase sample by s_a, the dwellings identified as respondents and non-respondents at the time of subsampling by s_{a1} and s_{a2}, respectively, and the subsample of non-respondents by s_2.

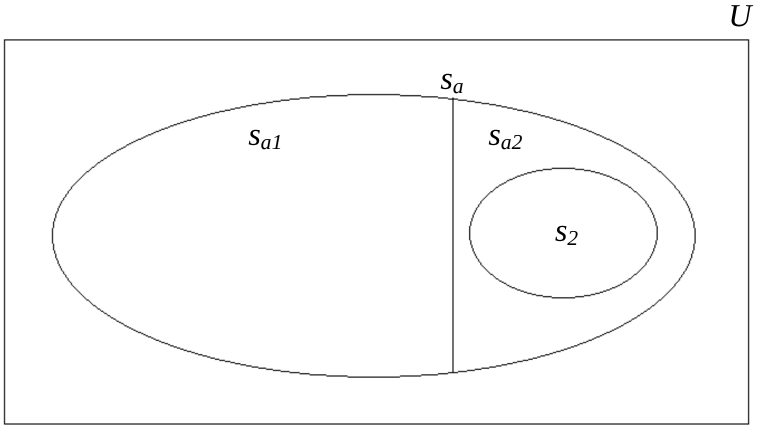

Figure 1: The NHS sample and follow-up subsample

In the Hansen and Hurwitz design, all the units of the subsample are assumed to respond. When this happens the following estimator of the total of the characteristic *y* can be used and is free of non-response bias:

$$\hat{t}_{y,\mathrm{HH}} = \sum_{k \in s_{a1}} \frac{y_k}{\pi_{ak}} + \sum_{k \in s_2} \frac{y_k}{\pi_{ak} \pi_{bk|s_{a2}}},$$

where π_{ak} is the probability of selection of unit *k* in s_a and $\pi_{bk|s_{a2}}$ is the probability of selection of unit *k* in the subsample s_2 given that s_{a2} is the group of non-respondents at the time of subsampling.

Full response in the subsample was not an assumption that could be made in the NHS. This means that it could not be expected that the non-response bias would be eliminated as in the traditional Hansen and Hurwitz setting and that the objective of subsampling in the NHS could not be to completely eliminate non-response bias. The objectives were rather to limit non-response bias, to control and concentrate follow-up costs and efforts and to target specific population groups.

At the time of subsampling, 37.5% of the in-scope households of the first-phase sample were eligible for subsampling (i.e., part of s_{a2}). The dwellings eligible to be subsampled were stratified geographically. CUs were used as the building blocks of the subsampling strata. The CUs were divided in two groups: those that had a homogeneous population and those that had a heterogeneous population. When the population of a stratum is heterogeneous the risk of non-response bias is more important than when it is homogeneous because there is a higher risk that respondents differ from non-respondents. To minimize this risk the subsampling fraction was chosen to be higher in heterogeneous strata. The CU was used as the subsampling stratum unless its size was too small. If that was the case, the CU was combined with a few others that were geographically close and similar in terms of heterogeneity. In each stratum, a systematic random subsample of dwellings was selected. The subsampling fraction was around 1/3.

In terms of estimation, the first phase weights were adjusted to take subsampling into account as in the Hansen and Hurwitz estimator. Thus the weights of the dwellings of s_{a1} were unchanged, the weights of the dwellings of s_{a2} not selected in the subsample s_2 were set to zero, and the weights of the dwellings that were part of the subsample were divided by the probability of selection in the subsample. The resulting weight will be called the double-expansion weight.

2.3 Adjustment for In-scope Households

Many dwellings that received a non-zero sampling weight do not house a household that is in-scope for the NHS. After the in-scope status of each dwelling is determined, the weight of out-of-scope dwellings is set to 0 so that they do not contribute to the NHS estimates. In practice, the dwellings with a weight of 0 after the adjustment for in-scope households do not contribute to the estimates because they are removed from the dissemination files. This is equivalent to using a domain indicator for in-scope status for every study variable of the survey. The rest of this paper is written under the assumption that such a domain indicator is part of every possible variable of study.

3. Further Adjustments Due to the Voluntary Nature of the NHS

As was expected, not every in-scope household in the NHS subsample responded to the survey during the follow-up. Among the in-scope households selected in the subsample, 43% responded. This departure from the Hansen and Hurwitz design can be measured using a weighted response rate to the survey. The weight that should be used is the double-expansion weight. If all the subsample had responded to the survey as assumed in Hansen and Hurwitz's theory, the weighted response rate would be 100% and there would be no non-response bias. The weighted response rate of the NHS is 77.2%. In comparison, the response rate of the 2006 Census long-form was 94%.

In fact, the weight adjustment corresponding to subsampling and described in section 2.2 is a form of non-response adjustment. Moreover, unlike the majority of non-response adjustments, it is free of any modeling assumptions and, when full response is obtained in the subsample, it is the only non-response adjustment needed to eliminate non-response bias. The weighted response rate of 77.2% of the NHS means that further non-response adjustments are necessary to compensate for the 22.8% weighted non-response. This section describes the estimation processes beyond the sampling weight adjustment of section 2.2 that were put in place to deal with the voluntary nature of the NHS.

3.1 Adjustment for Total Non-response in the Subsample

An adjustment had to be made to compensate for the household non-response observed in the subsample. The objectives of this adjustment were to minimize the non-response bias and to meet the overall goal of the NHS, which is to produce reliable estimates at various geographical levels (e.g., national, provincial, territorial and municipal).

3.1.1 Auxiliary information

In order to produce an adjustment that minimizes the non-response bias, it is crucial to have good auxiliary information known for both respondents and non-respondents. In the NHS, very detailed geographical information was available from the sampling frame. Furthermore, the census and the NHS were linked by design and the ten questions shared by the two surveys provided relevant auxiliary information for the adjustment. Before these data were used for the non-response adjustment, several steps were undertaken to obtain the most complete information possible:

1. Total non-response to the census was minimized by copying the responses of households that responded to the NHS but that did not respond to the census. Note that the non-response rate to the census was less than 3%.
2. The rest of the household non-response to the census was dealt with by whole household imputation (see Dolson, 2012). In this process, imputation for non-respondents was done using the data of the nearest geographic neighbour sharing the same household size which could be known or assigned prior to imputation.
3. For a given dwelling, when a response was obtained for both surveys and when it was verified that the same household provided all the answers, the two sets of response to the ten questions were compared and only the most complete set of information was used for both surveys.
4. Households were defined as in-scope for the NHS if they responded to the NHS or if, according to the census, they were in-scope. When an in-scope household did not respond to the NHS, the census information was copied to the NHS and used for the non-response adjustment.
5. In both surveys, item non-response to the ten questions was dealt with either deterministically or using the nearest-neighbour imputation methodology of the CANadian Census Edit and Imputation System (CANCEIS) (Liu *et al.*, 2011).

The responses to the ten census questions and the geographic information are very relevant because they describe many dimensions of the population that could explain the non-response mechanism, and because they are used to define a wide range of domains of interest in the NHS. However, the goal of the NHS is to measure characteristics of the Canadian population that were not measured by the census and there could still be an important residual non-response bias for those characteristics even if all the bias related to the census variables is removed. As a consequence, to enrich the information used for the non-response adjustment, several probabilistic linkages of pertinent 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 the years 1980 to 2010. Furthermore, some data known at an aggregate level was used for geographic areas corresponding to Indian reserves. For each household of such areas, it was identified which Indian band and tribe was predominant in its reserve.

3.1.2 Estimation options to deal with total non-response in the subsample

The first option that was considered to adjust for household non-response in the NHS was to treat this non-response as <u>item</u> non-response. It consists in copying the census data for NHS non-respondents and in imputing the missing information of the subsample non-respondents using the available auxiliary information. Donor imputation with CANCEIS was the only imputation method considered because it has been the traditional, and successful, imputation approach used for the Canadian census. Imputation was also appealing because it has been used to deal with both item and unit non-response in the previous censuses. Therefore, using CANCEIS gave an advantage in terms of both the systems and the expertise available. Moreover, as was already mentioned, the NHS aims at producing reliable estimates at various geographical levels and this option was attractive because it insured that there were a minimum number of responses for small geographies.

The Hansen and Hurwitz estimator, however, had to be adapted to take imputation into account. The following estimator makes use of the double-expansion weight:

$$\hat{t}_{y,\text{imp}} = \sum_{k \in s_{a1}} \frac{y_k}{\pi_{ak}} + \sum_{k \in s_{2r}} \frac{y_k}{\pi_{ak}\pi_{bk|s_{a2}}} + \sum_{k \in s_{2nr}} \frac{y_k^*}{\pi_{ak}\pi_{bk|s_{a2}}},$$

where s_{2r} and s_{2nr} are the sets of respondents and non-respondents of the subsample, respectively, and y_k^* are the imputed values. Another variant of this option was also studied where the missing data of every NHS non-respondent was imputed regardless of them being selected in the subsample or not. This variant was called *mass imputation*, while the former variant was called *partial imputation*. The estimator under mass imputation is

$$\hat{t}_{y,\text{ mass imp}} = \sum_{k \in s_{a1}} \frac{y_k}{\pi_{ak}} + \sum_{k \in s_{2r}} \frac{y_k}{\pi_{ak}} + \sum_{k \in s_{a2} - s_{2r}} \frac{y_k^*}{\pi_{ak}} \ .$$

The second option that was considered to adjust for household non-response to the NHS was to treat this non-response as unit non-response. With this option, the objective is to explain the mechanism leading to household non-response within the subsample and to increase the weight of the respondents of the subsample accordingly so that they also represent the non-respondents of the subsample. This is appealing because the auxiliary information is abundant and relevant to explain the non-response mechanism. It is also the typical strategy used to deal with unit non-response in voluntary surveys. This option was named *pure weighting* because it relies on weighting and not on imputation.

In terms of estimation, although no imputation is necessary, the double-expansion weight needs to be further adjusted. When the probability of response of each responding household of the subsample is estimated, the following estimator can be used:

$$\hat{t}_{y,\text{wgt}} = \sum_{k \in s_{a1}} \frac{y_k}{\pi_{ak}} + \sum_{k \in s_{2r}} \frac{y_k}{\pi_{ak}\pi_{bk|s_{a2}}\hat{p}_{k|s_2}}, \qquad (1)$$

where $\hat{p}_{k|s_2}$ is the estimated probability of response of unit *k* given that it is part of the subsample.

Those options were evaluated and compared using 2006 long-form data of the Toronto Census Metropolitan Area (CMA) by simulating the non-response and subsampling mechanisms (Verret *et al.*, 2011). Absolute errors were calculated as the absolute difference between the point estimator of the total of the option considered minus the Hansen and Hurwitz point estimator of the total. The mass imputation variant of the first option gave the largest absolute errors because knowledge of the subsampling mechanism was ignored and replaced by an imputation model, contrary to the partial imputation variant of the first option and to the pure weighting option. The partial imputation approach gave smaller errors than the pure weighting approach. The differences were very small at the CMA level, but at the sub-CMA level they were greater (the CMA was divided into close to a thousand sub-CMA geographical units for the analysis). Partial imputation was the first method implemented on a full scale in the NHS because of those simulation results, because it was the method with the largest effective sample size that took both phases of the design into account and because the systems to run partial imputation were ready while those for pure weighting were not. However, pure weighting was also implemented on a full scale in the NHS and was the estimation method used to deal with household non-response for reasons that will be given in the next section.

3.1.3 Implementing partial imputation

Partial imputation was implemented, using CANCEIS, in the following fashion. For each non-responding household of the subsample, a distance was measured between the non-respondent and each respondent of the subsample with the same household size. Additionally, households from Indian reserves were processed independently of households not living on Indian reserves. Each auxiliary variable contributed to the distance. For a given variable, if the value of the non-respondent and the value of the respondent were equal, then the contribution of the variable to the distance was zero; otherwise it was a positive number in the interval [0, 1], closer to 1 if the two values were very dissimilar. If the values of all variables were identical then the overall distance was zero. After each distance was calculated a list of the five nearest neighbours was created and a neighbour (i.e., a donor) was selected at random from the list to provide all the missing values of the non-responding household.

The relative weight given to each auxiliary variable in the distance measure has an impact on which donors are selected for imputation. Indeed, although the number of potential donors in the NHS is much larger than in most household surveys, it was not large enough to guarantee that every non-respondent would find a perfect donor (one that matches on every auxiliary variable). Moreover, it is extremely difficult to find the most appropriate weights using modelling because of the multivariate nature of the auxiliary data and of the data to be imputed. The most appropriate relative weights might vary from subject matter to subject matter. As a compromise, relative weights that gave similar unmatched rates for every auxiliary variable were sought. To achieve this, partial imputation was performed several times adjusting the weights each time.

It was found that the non-zero unmatched rates caused some improbable combinations of copied census data and imputed data at aggregate levels – significantly inflating the proportions of important outlier groups. For example, at the Canadian level, a lot more teenagers had PhD degrees than expected. This happened because too many teenage non-respondents had their data imputed using those of older respondents who had a PhD. Thus, the large number of responses in the NHS had a dual effect on partial imputation: it gave a better pool of potential donors than if the number of responses had been smaller, but it also made the inflation of outlier groups very evident at aggregate levels. The latter was because even a very small proportion of mismatched non-respondent/donor pairs for a key matching variable created a much greater number of outliers associated with the matching variable than were known to exist in the population.

Consequently, a second version of partial imputation was done to get rid of the outlier problems. The parameters were changed to allow the unmatched rates to be lower for auxiliary variables that were more critical according to subject matter analysts. Partial imputation was also done in two passes: some variables were imputed with the data of the nearest responding household, while the rest were imputed with the data of the nearest responding person. This gave better results, but outlier inflation was still found. For example, outliers were created between sex and occupation because the relative weight of sex in the distance measure gave a non-respondent/donor matching rate on this variable that was very high (well over 90%) but still not high enough. Partial imputation created, for example, far too many female plumbers and mechanics, and many male manicurists. Partial imputation was abandoned at this point because it seemed impossible to find values of the imputation parameters that would produce no critical outlier inflation problems within the tight production schedule.

3.1.4 Implementing pure weighting

The only option left to deal with household non-response in the NHS at that point was pure weighting and this was the final method that was implemented. Under this estimation approach, the problem of improbable combinations of census and non-census data found in partial imputation is absent because the only data that are used for estimation are those coming from questionnaires of the responding households. A standard pure weighting approach is to use the scores method in the subsample (Eltinge & Yansaneh, 1997). Under this approach, the probability of response of each respondent and non-respondent is first estimated with logistic regressions. Then, based on those probabilities and using clustering algorithms, response homogeneity groups are formed. Final probabilities of response are estimated in each group by the observed response rate in the group. Because of time issues, the scores approach could not be employed at the time pure weighting was adopted. If there had been enough time, many logistic regression models would have been constructed at geographical levels sufficiently small to get estimated probabilities of response that were well adapted to local response tendencies and sufficiently large to get sample sizes that would have provided good power and precision in the models. Although the scores method could not be implemented on a full scale in the NHS, the pure weighting approach used attempted to replicate it. As a first step, a few logistic regressions were done at the national level to judge the relative importance of the auxiliary variables in predicting the probability of response. As a second step, to get an adjustment that is good at the local level, nearest-neighbour searches, using CANCEIS, were done based on the relative importance of auxiliary variables found in the first step and the weight was transferred from each non-responding household of the subsample to its nearest responding neighbours in the subsample.

For the first step, one logistic regression was done for each household size. Households of size nine or more were grouped with households of size eight because there were too few of them to perform a logistic regression for each of their household size. An independent set of logistic regressions was done for households living on Indian reserves. A forward approach was used to include the variables one by one in the logistic regression models and the relative importance of each variable was measured by the increase in the R-square of the model. Every auxiliary variable that did not enter the model received a relative weight of half the weight of the last variable entering the model so that each variable considered would contribute to the non-response adjustment.

For the second step, nearest-neighbour searches were performed again and the results from the logistic regressions were used to assign the relative weights of each auxiliary variable in the distance function. Additionally, the number of donors in the best donors list was increased from five to twenty to mimic what is usually done in the clustering algorithm of the scores method. In this algorithm, the response homogeneity groups are formed so that they comprise a minimum number of responding units (e.g., twenty) to avoid extreme estimated probabilities of response and extreme non-response weight adjustments. The double-expansion weight of each non-responding household was transferred to each of its best twenty responding neighbours. To each of the twenty nearest neighbours in the list corresponds a distance to the non-responding household. Because the neighbours with the shortest distance are more similar to the non-respondent, the weight was transferred proportionally to the inverse of the distance. Moreover, each responding household might appear in the nearest-neighbour list of many non-respondents. This means a given respondent might receive a large amount of weight from

the various transfers. The total weight transferred to each responding household was capped to the observed 99th percentile to avoid extreme weight adjustments.
The estimator of a total corresponding to this weighting methodology, omitting the capping step, can be expressed in the following way. Let d_k represent the double-expansion weight:

$$d_k = \begin{cases} 1/\pi_{ak}, & \text{if } k \in s_{a1} \\ 1/\pi_{ak}\pi_{bk|s_{a2}}, & \text{if } k \in s_2 \\ 0, & \text{if } k \in s_{a2} - s_2. \end{cases}$$

Let $D(k,l)$ be the distance between households k and l, and η_l denote the set of 20 nearest responding neighbours of non-respondent l. Then the weight transferred from non-respondent l to respondent k is given by

$$\tau_{kl} = \begin{cases} d_l \dfrac{1/D(k,l)}{\sum_{i \in \eta_l} 1/D(i,l)}, & \text{if } k \in \eta_l \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, let $s_{2nr,k} = \{l \in s_{2nr} : k \in \eta_l\}$ be the set of non-respondents of the subsample who have respondent k in their 20 nearest-neighbours list. Then the total weight transferred to respondent k is given by

$$\tau_{k\bullet} = \sum_{l \in s_{2nr,k}} \tau_{kl} = \sum_{l \in s_{2nr,k}} d_l \frac{1/D(k,l)}{\sum_{i \in \eta_l} 1/D(i,l)}.$$

The corresponding point estimator of a total is given by

$$\hat{t}_{y,\text{wgt}} = \sum_{k \in s_{a1}} d_k y_k + \sum_{k \in s_{2r}} (d_k + \tau_{k\bullet}) y_k \ .$$

3.1.5 Lessons learned

In the search for the most appropriate estimation method to compensate for household non-response in the NHS and the implementation of those methods, some important lessons were learned. With regards to mass imputation and partial imputation, it was found in the simulations that the second phase of sampling should not be ignored. Partial imputation takes this phase into account and mass imputation does not. Knowledge of the second phase design should not be replaced by an imputation model.

With regards to using a whole-household and nearest-neighbour imputation approach to deal with household non-response (i.e., impute all the missing values of a given non-respondent by those of one responding household), the probability of getting high rates of outliers is higher when a lot of information is known about the non-respondent and when a lot of information is imputed. To minimize this risk, one should have one or the other but not both. This is the case of the whole household imputation approach used to deal with the household non-response in the census, where only the geography and household size are known. Note that the importance of this problem also increases with the extent of the observed total non-response. This lesson holds even for very large scale surveys like the NHS for which a large pool of respondents is available. This could be seen as a curse of dimensionality. The pool might be large enough to find close neighbours when a few auxiliary variables are used in the distance function (e.g., as is seen in the regular census and NHS item non-response imputation processes), but as the number of auxiliary variables increases, the pool size to get close neighbours needs to get larger and larger and possibly at an exponential rate.

3.2 Adjustment for Total Non-response among the Households not Eligible for the Subsampling

In the Hansen and Hurwitz setting, every unit of s_{a1} is assumed to have responded when the subsample is selected. Unfortunately in the NHS this was not the case. Among the in-scope households selected at the first phase, 61.5% were part of s_{a1} and 1.4% were non-respondents in s_{a1}. Two main reasons explain this. Firstly, at the time subsampling was done, full analysis of the questionnaires received was not completed. The response status was based on field information. A given household could have been identified in the field as having responded, but the analysis of the questionnaire could reveal the opposite. For example, the household could have returned an empty questionnaire. Secondly, some addresses were added to the sampling frame just before collection started and because of operational issues it was not possible to provide NHS questionnaires to those addresses. A number of those dwellings housed a household that was in-scope for the NHS. Those households thus had no opportunity to respond to the survey.

The adjustment for non-response in s_{a1} is similar to the adjustment used to deal with non-response in the subsample. The pure weighting methodology described previously was applied with the difference that the nearest neighbours could be found in the subsample or in s_{a1}. The estimator of a total can now be expressed in the following way (still omitting the capping of the weight transfer step):

$$\hat{t}_{y,\text{wgt}} = \sum_{k \in s_{a1r}} \left(d_k + \tau_{a1k\bullet} \right) y_k + \sum_{k \in s_{2r}} \left(d_k + \tau_{a1k\bullet} + \tau_{2k\bullet} \right) y_k ,$$

where s_{a1r} and s_{a1nr} are the sets of respondents and of non-respondents of s_{a1} respectively, $\tau_{a1k\bullet} = \sum_{l \in s_{a1nr,k}} \tau_{kl}$, $\tau_{2k\bullet} = \sum_{l \in s_{2nr,k}} \tau_{kl}$ and $s_{a1nr,k} = \left\{ l \in s_{a1nr} : k \in \eta_l \right\}$.

3.3 Adjustment for Surprise Respondents

One last adjustment deals with the voluntary nature of the NHS. This time it was needed because of extra responses to the survey. Some households had not responded to the survey at the time of subsampling, were eligible for subsampling, were not selected in the subsample but chose to respond nonetheless. Those households were called the surprise respondents. They represent 1.8% of the in-scope households of the first phase sample and the Hansen and Hurwitz design do not cover them.

The easiest way to deal with surprise respondents estimation-wise is to ignore their response and to not consider them in the estimates (i.e., keep their weight at zero). This is the approach that should minimize the non-response bias because surprise non-respondents are units that are more inclined to respond. However, it is not acceptable to effectively eliminate valid responses to the survey. Various estimation strategies could be employed to assign weights greater than zero to surprise respondents. One natural approach would be (ignoring the non-respondents of s_{a1} for this particular example) to replace $\pi_{bk|s_{a2}}\hat{p}_{k|s_2}$ by $\pi_{bk|s_{a2}}\hat{p}_{k|s_2}+\left(1-\pi_{bk|s_{a2}}\right)\hat{q}_{k|s_{a2}-s_2}$ in (1), where $\hat{q}_{k|s_{a2}-s_2}$ is the estimated probability of response of household *k* given it is eligible for subsampling and it is not selected in the subsample. That is, replacing the estimated probability of being in the subsample and of responding by the overall estimated probability of responding. Under this approach, every surprise respondent receives a weight greater or equal to the first phase weight. This strategy involves modeling in some fashion the probability of response of households of the subsample and the probability of response of households eligible for subsampling that were not subsampled.

Instead of using this approach, the weight of each surprise respondent was increased to 1 and the weight of respondents in the subsample was reduced accordingly. The reason for choosing this value is that it is the weight closest to 0 that guarantees each surprise respondent is at least self-representative. This was done using again a nearest-neighbour approach. The 20 nearest responding neighbours in the subsample of each surprise respondent were found and a weight of 0.05 was transferred from each to the surprise respondent. Let s_{surprise} and $s_{\text{surprise},k}$ denote the set of surprise respondents and the set of surprise respondents who have respondent *k* as one of their 20 nearest neighbours respectively. The estimator of a total, omitting the cap on the weight transfer, is then given by

$$\hat{t}_{y,\text{wgt}} = \sum_{k\in s_{a1r}} \left(d_k+\tau_{a1k\cdot}\right)y_k + \sum_{k\in s_{2r}}\left(d_k+\tau_{a1k\cdot}+\tau_{2k\cdot}-\sum_{s_{\text{surprise},k}} 1/20\right)y_k + \sum_{k\in s_{\text{surprise}}} y_k .$$

4. Weight Calibration

The last step of the point estimation process is weight calibration to census totals. The goals of calibration are to obtain coherence between NHS estimated totals and corresponding census totals, as well as a reduction of the variance. To perform calibration, the country is first split geographically into Weight Calibration Areas (WCAs). Calibration constraints selection and calibration on the chosen constraints is then done WCA by WCA.

WCAs are formed by combining adjacent dissemination areas (DA). While the CUs are used to manage collection activities, DAs are used for dissemination purposes. A DA contains between 400 and 700 persons. A WCA represents on average 2300 dwellings and 5700 persons. There were 5884 WCAs created for the 2011 NHS. The country was split into WCAs to enable the iteration of constraint selection and calibration, and to insure the calculations, such as the matrix inversions required for calibration and variance estimation, were doable systems-wise. In the construction of WCAs, CD boundaries are respected and boundaries of lower levels of geography (census sub-divisions and the census tracts) are respected as much as possible. The WCAs do not vary much in size so that a common list of potential calibration constraints can be used for all WCAs.

There are many calibration constraints that can be used because of the ten questions in common between the census and the NHS. The constraints that were considered for selection within each WCA were:

- Number of households and number of persons in the WCA and in each of the DAs of the WCA

- Sex and sex crossed with indicator of age less than 15 years old
- 5-year age groups up to 75 years old
- Marital status
- Number of census families with or without children
- Census and economic family indicators
- Number of households living in semi-detached houses and number of households living in apartment buildings of less than 5 stories
- Household size
- Common combinations of mother tongue, official languages and home language.

In each WCA, the calibration constraint selection process was performed in the following way. The constraints were divided into three groups. First, some constraints were identified as mandatory in every WCA because it was required to have the NHS estimated total equal to the census total at the national level. The mandatory constraints were number of households and number of persons in the WCA. Secondly, in a given WCA, there may have been constraints that should have been avoided. This was the case with every constraint that had less than 30 responses because calibrating on those constraints might have hurt the precision of the estimates instead of improving it. These will be called the imprecise constraints. The rest of the constraints might have been selected for calibration. These will be called the available constraints.

The constraints were chosen in a forward fashion inspired by Verret and Kevins (2010). The two mandatory constraints were included at the first step. In the following step, each of the non-selected available constraint was evaluated and could have been discarded for the rest of the selection process. The best non-discarded constraint was included (if this set was not empty). This was repeated until all available constraints were either included or discarded. An available constraint was discarded if its inclusion would have been equivalent to calibrating on an imprecise constraint or on a combination of included constraints. For example, it might have been desirable to avoid calibrating on number of males because their responses were too few, while total number of persons was mandatory and number of females was available. In this case, exclusion of number of males would not be sufficient to avoid calibrating on this constraint because inclusion of the other two constraints would have been equivalent to calibrating on number of males by construction. An example of the second problem is to have total number of persons mandatory and number of males and number of females available. In this case, only number of males or number of females would have to be chosen but not both, to avoid redundancy.

To perform this evaluation of the available constraints, a non-redundant set of the imprecise constraints was first created. Let the calibration micro-data information of the NHS respondents be partitioned into Z^0, the mandatory constraints, Z^m, the imprecise constraints and Z, the available constraints. In these matrices rows correspond to respondents and columns to constraints. Let X^m represent the matrix of non-redundant imprecise constraints that is sought. This matrix was first set equal to Z^0. Then columns of Z^m were either added one by one to X^m in a forward fashion or discarded from further selection. At each step, a regression was performed for each column of Z^m that was neither discarded nor added to X^m at a previous step. Each column was regressed on X^m. If the R-square was too high then the constraint was discarded because it is redundant. The remaining column (if any) with the minimum R-square was added to X^m.

This was repeated until all columns of Z^m were either discarded or added. After this was done, the columns of Z^0 were removed from X^m.

The available constraints were then evaluated. Let *X* denote the matrix of chosen constraints. This matrix was set equal to Z^0 at the first step of selection. At each following step, a regression was done for each column of *Z* that was neither discarded nor selected. Each column was regressed on *X* augmented by X^m. If the R-square was high then the constraint was discarded. The remaining constraint (if any) with the minimum R-square was selected and its column was added to *X*. The process ended when every column of *Z* was either discarded or selected.

Statistics Canada's StatMx software was used to perform calibration on the constraints that were selected WCA by WCA. It finds the set of calibrated weights w_k such that $\sum_{k\in s_r} c_k\left(w_k - a_k\right)^2 \big/ a_k$ is minimized, subject to the calibration constraints $\sum_{k\in s_r} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k$ and the constraint $l_k \le w_k \le u_k$ for $k \in s_r$, where s_r is the set of respondents, c_k are chosen constants, a_k is the uncalibrated weight and l_k and u_k are lower and upper bounds on the calibrated weights. In the NHS, $s_r = s_{a1r} \cup s_{2r} \cup s_{\text{surprise}}$, $c_k = 1$, a_k is the weight after the adjustments of section 3, $l_k = 1$ and $u_k = 100$.

5. Variance Estimation

For variance estimation, it was desirable to use as much as possible a design-based approach for simplicity. To do this, the household non-response in s_{a1} and in the subsample were treated as a phase of sampling. Furthermore, surprise respondents were ignored in the variance estimation process, also for simplicity. Since sampling fractions were large, both phases of sampling and the non-response phase had to be taken into account in variance estimation. The stratified systematic designs of the sampling phases were approximated by stratified simple random sampling without replacement designs. This approximation is reasonable because the design strata are small and could thus be supposed to be homogeneous. The approximation was also needed to get explicit variance estimators. Moreover, the household non-response that was observed was supposed to have originated from a Poisson phase happening after the second phase of sampling. The probability of response of respondents was estimated by the inverse of the non-response weight adjustment and was assumed known for variance estimation purposes. It was equal to $d_k \big/ \left(d_k + \tau'_{a1k\bullet}\right)$ for respondents of s_{a1} and to $d_k \big/ \left(d_k + \tau'_{a1k\bullet} + \tau'_{2k\bullet}\right)$ for respondents of the subsample, where $\tau'_{a1k\bullet}$ and $\tau'_{2k\bullet}$ are the capped version of $\tau_{a1k\bullet}$ and $\tau_{2k\bullet}$ respectively. For example, a respondent that had its weight doubled by the weight transfer had an estimated probability of response of 50%.

Variance estimators were developed by doing classical developments, expanding the Hansen & Hurwitz variance estimator to the case where Poisson sampling is done. The Hansen & Hurwitz estimator can be seen as a two-phase variance estimator when s_{a1} is seen as a take-all second phase stratum. The resulting estimator thus has the form of a three-phase variance estimator. Taylor linearization was used to take calibration into account.

StatMx was used as well to estimate the variance. However it had to be slightly adapted because StatMx estimates the variance for one or two phase designs. The modification is based on an identity similar to one derived in Lundström & Särndal (1999). Under a three-phase design, let

- s_a , s_b and s denote the first, second and third phase samples;
- π_{ak} , π_{bk} and π_k denote the corresponding first-order inclusion probabilities of unit *k*;
- π_{akl} , π_{bkl} and π_{kl} denote the corresponding second-order inclusion probabilities of units *k* and *l*;
- and Δ_{akl} , Δ_{bkl} and Δ_{kl} denote the corresponding covariances between the inclusion indicators of units *k* and *l*.

In the absence of calibration, the variance estimator of the estimator of a total of *y* is given by

$$v\left(\hat{t}_y\right)=\sum_s\sum_s\frac{\Delta_{akl}}{\pi_{akl}\pi_{bkl}\pi_{kl}}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}}+\sum_s\sum_s\frac{\Delta_{bkl}}{\pi_{bkl}\pi_{kl}}\frac{y_k}{\pi_{ak}\pi_{bk}}\frac{y_l}{\pi_{al}\pi_{bl}}+\sum_s\sum_s\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_{ak}\pi_{bk}\pi_k}\frac{y_l}{\pi_{al}\pi_{bl}\pi_l}.$$

When the third phase is Poisson then

$$\pi_{kl}=\begin{cases}\pi_k, & \text{if } k=l\\ \pi_k\pi_l, & \text{otherwise}\end{cases} \text{ and } \Delta_{kl}=\begin{cases}\pi_k\left(1-\pi_k\right), & \text{if } k=l\\ 0, & \text{otherwise.}\end{cases}$$

Let

$$z_k=\begin{cases}y_k/\pi_k\,, & \text{if } k\in s\\ 0, & \text{if } k\in s_b-s.\end{cases} \tag{2}$$

The estimator of the variance can then be rewritten in terms of *z* as

$$v\left(\hat{t}_y\right)=\sum_{s_b}\sum_{s_b}\frac{\Delta_{akl}}{\pi_{akl}\pi_{bkl}}\frac{z_k}{\pi_{ak}}\frac{z_l}{\pi_{al}}+\sum_{s_b}\sum_{s_b}\frac{\Delta_{bkl}}{\pi_{bkl}}\frac{z_k}{\pi_{ak}\pi_{bk}}\frac{z_l}{\pi_{al}\pi_{bl}}+\sum_{s_b}\pi_{ak}\pi_{bk}\left(1-\pi_k\right)\left(\frac{z_k}{\pi_{ak}\pi_{bk}}\right)^2. \tag{3}$$

The first two terms are those of the two-phase variance estimator of $\hat{t}_z$. To estimate the variance of the estimator of the total of *y*, the second phase sample $\left(s_{a1}\cup s_2\right)$ can be submitted to StatMx with values of *z* instead of those of *y* and the third term of the equation can be programmed within a special edition of StatMx and added to the total estimated variance.

In the presence of calibration, estimator (3) should be used with y_k replaced by $g_{ks}e_{ks}$ in the definition of z_k given by (2), where g_{ks} is the calibration weight adjustment or the *g*-weight, $e_{ks}=y_k-\mathbf{x}_k\hat{\mathbf{B}}$ and $\hat{\mathbf{B}}=\left(\sum_s\frac{\mathbf{x}_k\mathbf{x}_k^T}{\pi_{ak}\pi_{bk}\pi_k c_k}\right)^{-1}\sum_s\frac{\mathbf{x}_k y_k^T}{\pi_{ak}\pi_{bk}\pi_k c_k}$. It is possible to use StatMx to calculate this variance, provided the third term of (3) is added to the total variance in the program. This time the second phase sample can be submitted to StatMx with the g-weights obtained from the point estimation run of StatMx, $c_k = 1/\pi_k$ and with y_k and $\mathbf{x}_k$ replaced by

$$y'_k = \begin{cases} y_k/\pi_k, & \text{if } k \in s \\ 0, & \text{if } k \in s_b - s, \end{cases} \quad \text{and } \mathbf{x}'_k = \begin{cases} \mathbf{x}_k/\pi_k, & \text{if } k \in s \\ 0, & \text{if } k \in s_b - s. \end{cases}$$

6. Conclusion

The NHS is the largest voluntary household survey conducted by Statistics Canada. Its voluntary nature required many sampling, collection and estimation innovations. This paper focused on the innovations in the estimation methodologies. Two-phase design weights were calculated to match the sampling design. Before and during production, a lot of research on adjustments for household non-response was achieved and important lessons were learned. Pure weighting was preferred to whole household imputation in the end. However, the results of the nearest-household-neighbour search obtained using CANCEIS drove the non-response weight adjustment. Weight calibration to known census totals was done after selection of constraints. The selection involved identifying imprecise constraints and avoiding calibration on them as well as on redundant constraints. Finally, three-phase variance estimators were developed and used. The Taylor linearization technique was used to deal with calibration in the variance estimators.

References

Dolson, D. (2012), “Errors of Non-observation: Dwelling Non-response and Coverage Error in Traditional Census,” in *JSM Proceedings*, Survey Research Methods Section. San Diego, CA: American Statistical Association. 3395-3409.

Eltinge, J. L., and Yansaneh, I. S. (1997), “Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey,” *Survey Methodology,* 23, 1, 33-40.

Hansen, M. H., and Hurwitz, W. N. (1946), “The Problem of Non-Response in Sample Surveys,” *Journal of the American Statistical Association*, 41, 517-429.

Liu, W., Crowe, S., and Alavi, A. (2011), “How CANCEIS Does Edit & Imputation, and Can it be Standardized for more General Use?,” in *2011 International Methodology Symposium*. Ottawa, Canada: Statistics Canada.

Lundström, S., and Särndal, C.-E. (1999), “Calibration as a Standard Method for Treatment of Nonresponse,” *Journal of Official Statistics*, 15, 305-327.

Verret, F., and Kevins, C. (2010), “Calage aux Marges des Poids d’Enquêtes à Plan Complexe pour le Refus à l’Appariement,” in *SSC Annual Meeting Proceedings*, Survey Methods Section. Quebec, Canada: Statistical Society of Canada.

Verret, F., Bankier, M., Benjamin, W., and Hayden, L. (2011), “The Estimation Strategy of the National Household Survey (NHS),” in *International Total Survey Error Workshop 2011 Presentations*. Quebec, Canada: National Institute of Statistical Sciences.