

# CREATION OF A NEW LONGITUDINAL WEIGHT FOR THE CANADIAN NATIONAL POPULATION HEALTH SURVEY: PROVIDING DATA USERS WITH GREATER ANALYTICAL FLEXIBILITY

François Brisebois and Patrice Mathieu<sup>1</sup>

## ABSTRACT

Statistics Canada's National Population Health Survey (NPHS) is a longitudinal survey designed to collect information on the health of the Canadian population and related socio-demographic characteristics. For NPHS Cycle 4, a new longitudinal file was added to the existing set of files produced by the survey. This new file, the C1-C4 Full file, consists of the subset of all panel members that provided a full response to at least Cycles 1 and 4. This paper outlines the weighting strategy used to compute the weight associated with this new file.

KEY WORDS : Longitudinal survey, Total nonresponse, Weighting,

## RÉSUMÉ

L'Enquête nationale sur la santé de la population de Statistique Canada est une enquête longitudinale conçue pour recueillir de l'information sur la santé de la population canadienne ainsi que des renseignements socio-démographiques connexes. Lors du cycle 4 de l'ENSP, un nouveau fichier longitudinal a été ajouté à la liste de fichiers produits par l'enquête. Ce nouveau fichier, le fichier complet C1-C4, contient le sous-ensemble des membres du panel ayant fourni une réponse complète au moins aux cycles 1 et 4. Cet article présente un aperçu de la stratégie de pondération utilisée pour calculer le poids associé à ce nouveau fichier.

MOTS-CLÉS: Enquête longitudinale, pondération, non-réponse totale

## 1. INTRODUCTION

Statistics Canada's National Population Health Survey (NPHS) is a longitudinal survey designed to collect information on the health of the Canadian population and related socio-demographic characteristics. In 1994, over 20,000 households were contacted to provide the NPHS with its first cycle of data. General socio-demographic information as well as basic health questions were asked to each household member, followed by a more in-depth interview conducted with one randomly selected person. The information collected from the entire household would be used only for cross-sectional purposes, while the information obtained from the selected person would be used for both cross-sectional and longitudinal purposes. From all contacted households, a total of 17,276 selected people met minimum requirements to form the NPHS longitudinal panel and are recontacted every second year for a period of 20 years, that is, a total of ten cycles. Information collected from the panel members through all these years allows analysts to study and better understand the dynamic process of health. Readers are invited to consult Tambay and Catlin (1995) for more details about the design of the NPHS.

For NPHS Cycle 4, a new longitudinal file was added to the existing set of files produced by the survey. This new file, the C1-C4 Full file, consists of the subset of all panel members that provided a full response to at least Cycles 1 and 4. The main purpose of this paper is to present the strategy used to calculate the sampling weights assigned to the subset of panel members included in the Full C1-C4 file.

## 2. DEFINITION OF NPHS LONGITUDINAL FILES

As any other survey, the NPHS faces two types of nonresponse: total nonresponse and item nonresponse. Total nonresponse, or unit nonresponse, occurs when the person selected fails to respond to the survey. Figure 1 shows all patterns of unit response/nonresponse that occurred over the four first cycles of NPHS. The notation used indicates the cumulative pattern of unit response/nonresponse observed at each cycle. A response to a cycle is denoted as '1' in the

---

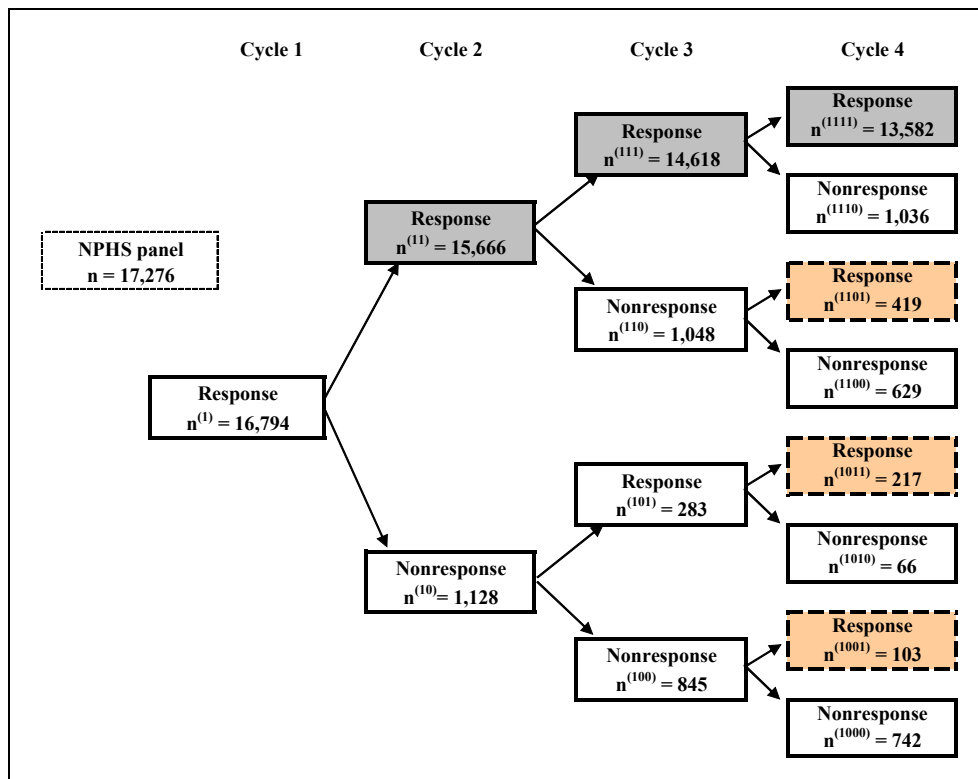
<sup>1</sup> François Brisebois(francois.brisebois@statcan.ca) and Patrice Mathieu(patrice.mathieu@statcan.ca), Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6

pattern, while a nonresponse is denoted as '0'. For example, panel members who have responded to all of the first four cycles of NPHS are labelled as '1111'. Sample sizes for each pattern are also reported in the figure.

Item nonresponse, also referred as partial nonresponse, refers to the situation in which a unit response is obtained, but the respondent does not answer all of the questions. For NPHS, item nonresponse is not imputed, but reported in the data files using a standard codification. The task of properly handling item nonresponse in analyses is left to the data users and is not discussed in this paper.

Once a given cycle is completed and its data are processed, a series of longitudinal data files is created. The following subsections describe the two data files traditionally created after each NPHS cycle (the Square file and the Full file), as well as the newest data file created at Cycle 4 : the C1-C4 Full file.

**Figure 1: Patterns of response/nonresponse for the first four cycles of NPHS**



## 2.1. The Square file

The *Square file* contains one record for each of the 17,276 panel members. All variables collected or derived in each cycle since Cycle 1 are included. After each cycle, the newly collected data are appended to the previous cycle's file. All variables from a cycle where a panel member did not respond are coded as missing. Similarly, if a panel member is deceased, all variables are set to missing for that cycle, and will be in all subsequent cycles as well. One important note from an analytical point of view is that although the deceased are coded with missing values, they still prove useful in morbidity analyses, and therefore are considered as respondents.

The richness of the data contained in the Square file definitely has a strong potential for analysis. However, users must deal with incomplete data, especially those resulting from total nonresponse. Although some methods can be used in analysis to overcome this obstacle, a second file is also provided to data users and has proven to be more practical from this point of view. This file is named the *Full file*.

## 2.2. The Full file

For the purpose of panel analyses, only individuals who responded to all cycles are generally of interest. Consequently, the Full file was created by including only records of panel members who provided a full response in all cycles (which includes panel members who are deceased). The file contains all variables accumulated since Cycle 1. Therefore, the

Full file can be seen as a subset of records from the Square file, which excludes all records with at least one occurrence of total nonresponse. At each cycle, these records are reweighted to take into account unit nonresponse from the current cycle. The weighting of the Full file is discussed in section 3.2. The groups of records that are part of the Full file at each cycle are displayed in Figure 1 as shaded boxes with plain borders. The Full file contained 15,666 records after Cycle 2 ( $n^{(11)}$ ), 14,618 records Cycle 3 ( $n^{(111)}$ ), and more recently, 13,582 records after Cycle 4 ( $n^{(1111)}$ ).

### 2.3. Creation of the Full C1-C4 subset

The Full file provides users with a more practical database than the Square file since it eliminates the need for special treatments of total nonresponse in analyses. However, as cycles go by, the Full file undergoes a substantial loss in sample size, and consequently, a substantial loss in precision occurs for analyses using this file. Moreover, although only panel members who responded to all cycles are included in the Full file, all other members are still recontacted every cycle. In cases where panel members excluded from the Full would respond at a subsequent cycle, their data would remain practically unused (although these data would appear in the Square file, so we should rather say they would be underused). For example, 739 panel members who were not part of the Cycle 4 Full file had responded to Cycle 4 (corresponding to the shaded cells with dotted borders in Figure 1). It was in great part based on these observations that a new file was created : the *Full C1-C4 file*.

Therefore, the Full C1-C4 file contains all panel members who responded to at least cycles 1 and 4, regardless of the status of their cycle 2 and 3 interviews. For users who want to focus their analyses on Cycle 1 and 4 variables, this file presents the advantage of considerably increasing the available sample size compared to the Full file. However, if Cycle 2 and 3 variables are also of interest, users should settle for the Full file unless they are willing to deal with incomplete data at these two cycles.

The end result is that on top of the 13,582 records ( $n^{(1111)}$ ) in the Full file, 739 extra records ( $n^{(1101)} + n^{(1011)} + n^{(1001)}$ ) are added for the creation of the Full C1-C4, which gives a total sample size of 14,321 (noted as  $n^{(1..1)}$ ). Since this is, like the Full file, a subset of the panel, an adjusted sampling weight is necessary to account for the potential bias of dropping some records in the creation of this new file. The strategy used to compute the sampling weight is discussed in section 3.

## 3. WEIGHTING OF THE FULL C1-C4 FILE

As described by Gelman and Carlin (2002), the essential idea of sample survey weighting is to correct for known differences between sample and population, whether these discrepancies arise from sampling fluctuation, nonresponse, frame errors, or other sources. The sampling weights can therefore be used to compute estimates that are approximately unbiased for the population surveyed.

This section presents the weighting strategy used to calculate the sampling weight for the Full C1-C4 file. First, sections 3.1 and 3.2 give some details about the weighting strategy used for the Square and Full files, which will help to better understand the strategy used for the Full C1-C4, presented in section 3.3.

### 3.1. Weighting of the Square file

For longitudinal surveys, the weighting of the first cycle is identical to computing the weight of a cross-sectional survey. Since the Square file includes all panel members interviewed in Cycle 1, its weighting was done using standard steps used in cross-sectional survey weighting. First, an initial weight defined as the inverse probability of selection in the sample was calculated. Next, the initial weight was adjusted in order to compensate for the nonresponse observed in Cycle 1. The adjustment was applied using a weighting class method where respondents and nonrespondents were classified by a number of variables thought to be informative of nonresponse. The main assumption of this approach is that respondents and nonrespondents within each class share the same propensity to respond to the survey. The limited sample design characteristics available for respondents and nonrespondents alike, were used to define membership in the adjustment classes. The inverse weighted response rate computed within each class was the factor used to adjust the initial weight of panel members. Finally, a poststratification was applied to ensure the weights are consistent with Census-based population estimates for 1994, the survey reference year. Since the composition of the Square file always remains the same from one cycle to the other (only newly collected variables are appended to the file), the sampling weights should remain exactly the same (unless population estimates are revised).

### 3.2. Weighting of the Full file

The Full file was first created for Cycle 2 and as described in section 2.2, contains only records of panel members who gave a full response at all cycles, up to and including the present cycle. Therefore, the Full file of each cycle consists of a subset of the previous cycle's Full file, and this is reflected in the weighting strategy. Indeed, the starting point for the weighting of one cycle is the previous cycle's Full file sampling weight (from which the poststratification was removed). Using adjustment classes, a weight adjustment is then performed to account for the previous cycle Full file respondents lost at the current cycle. Compared to the weighting of the Square file, where only limited information was available for both respondents and nonrespondents, much richer information is now available to define the classes. Panel members that are part of the Full file of one cycle all provided a full interview in previous cycles, and therefore all have a large number of variables available. Finally, as done for the Square file weight, a poststratification is applied using 1994 Census-based population estimates.

### 3.3. Weighting of the Full C1-C4 subset

For longitudinal files that include total nonresponses (as is the case for the C1-C4 Full file), the longitudinal weighting is usually done relying solely on Cycle 1 variables, since they are the only ones available from all panel members. Due to the varying amount of information available from one panel member to another, it is indeed much simpler to use such an approach. However, this approach does not benefit from the fact that variables measured at intermediate cycles (in our case Cycles 2 and 3) are available for a large portion of the panel members, and therefore could potentially be used to refine the weighting strategy. Moreover, since there is already an existing weight calculated for the Full file, it would be desirable to produce a C1-C4 Full weight that is consistent with the existing Full weight, or more precisely that they both should lead to similar estimates. Consequently, the weighting strategy was developed with two goals: i) being consistent with the Full weight, and ii) taking advantage of the fact that we have information available at other cycles than Cycle 1 for a large portion of the panel members.

Section 3.3.1 presents the notation used to calculate the Full C1-C4 weight. Section 3.3.2 describes the weighting strategy traditionally used in longitudinal surveys, and section 3.3.3 presents the proposed approach. Results of comparisons between the two methods are presented in 3.3.4 to evaluate what is gained from using the proposed approach.

#### 3.3.1 Notation

Let us denote :

WS : sampling weight calculated for the 17,276 panel members in the Square file

WF2 : sampling weight calculated for the 15,666 panel members part of the Cycle 2 Full file

WF4 : sampling weight calculated for the 13,582 panel members part of the Cycle 4 Full file

WE4 : sampling weight calculated for the 14,321 panel members part of the Cycle 4 C1-C4 Full file

$i$  : index identifying a panel member

$c_2$  : index identifying nonresponse cells based on Cycle 1 variables, and used for the weighting of the Cycle 2 Full file

$c_3$  : index identifying nonresponse cells based on Cycle 2 variables, and used for the weighting of the Cycle 3 Full file

The notation defined in section 2 identifying the different subsets of the panel according to their longitudinal pattern of response, is also used. Note that a dot (.) in the pattern indicates that both respondents and nonrespondents are included for that specific cycle.

#### 3.3.2 Weighting method #1: Traditional approach

The weighting of a longitudinal file that includes records showing some total nonresponse (such as our C1-C4 Full file) is traditionally done using a relatively simple approach. The longitudinal weight available for all panel members (in our case WS, the Square weight) is used as the starting point. Weights of records dropped from the longitudinal file created are redistributed to records remaining in the file using weighting classes formed using Cycle 1 variables (or even simpler, using sample design variables), which are available from all panel members. Finally, a poststratification is applied to create the desired longitudinal weight. This approach is considered simple in the sense that nonresponse cases are all treated at once, using only one set of adjustment classes.

### 3.3.3 Weighting method #2: Proposed approach

Since one of the goals is to obtain a weight that is consistent with the Cycle 4 Full file's weight, the latter was used as the starting point. Therefore, records that are part of the C1-C4 Full file but not the Full file do not have a starting weight. The general idea of the proposed approach for these records is to recuperate the latest weight that was assigned during the course of the survey, and then apply some adjustments. For the (1101) group, this corresponds to the Cycle 2 Full weight, while for the (1011) and (1001) groups, it corresponds to the Square weight. The necessary adjustments to these weights are done in two steps.

The first step intends to redistribute a portion of Cycle 4 Full file weight to the group of records lost from the Full file at Cycle 3, but recuperated by the C1-C4 Full file (the (1101) group). This portion consists of the weighted proportion the (1101) group represented among both (1111) and (1101) groups (denoted as (11.1) in equation (1)) in the Cycle 2 Full file. This proportion is computed within each adjustment class used in the weighting of Cycle 3 Full file (where the weight of the (1101) was last used), and is noted as follows:

$$P_{c_3}^{(1101)} = \frac{\sum_{i \in n^{(1101)}} WF2_{c_3 i}}{\sum_{i \in n^{(11.1)}} WF2_{c_3 i}} \quad (1)$$

The Cycle 2 Full weight used as the initial weight for the (1101) group is then boosted such as its sum, after adjustment, represents  $P_{c_3}^{(1101)} \times 100$  percent of the total weight resulting from this first step. To achieve this, Cycles 2 and 4 Full weights, as well as the estimated proportion calculated in (1) are combined for the (1101) group, as follows:

$$WE4'_{c_3 i} = WF2_{c_3 i} \times P_{c_3}^{(1101)} \left( \frac{\sum_{i \in n^{(1111)}} WF4_{c_3 i}}{\sum_{i \in n^{(1101)}} WF2_{c_3 i}} \right) \quad (2)$$

As for the (1111), their Cycle 4 Full weight must be decreased by the same proportion to respect the overall sum of weight. The equation used for this group, as well as the simplified equation for the (1101) group, is presented in the equation (3) below.

$$WE4'_{c_3 i} = \begin{cases} WF2_{c_3 i} \times \frac{\sum_{i \in n^{(1111)}} WF4_{c_3 i}}{\sum_{i \in n^{(11.1)}} WF2_{c_3 i}} & i \in n^{(1101)} \\ WF4_{c_3 i} \times \left[ 1 - P_{c_3}^{(1101)} \right] & i \in n^{(1111)} \end{cases} \quad (3)$$

The next step aimed to redistribute a portion of the weight derived in (3), from the (1101) and (1111) groups to the remaining (1011) and (1001) groups. In this case, the portion consisted of the weighted proportion the (1011) and (1001) groups represented in the Square file, among records to be included in the Full C1-C4 file (denoted as (1..1) in equation (4)). This proportion is computed within each adjustment class used for the weighting of the Cycle 2 Full file, and is expressed as follows:

$$P_{c_2}^{(10.1)} = \frac{\sum_{i \in n^{(10.1)}} WS_{c_2 i}}{\sum_{i \in n^{(1..1)}} WS_{c_2 i}} \quad (4)$$

The Square weight used as the initial weight for both (1011) and (1001) groups is boosted such as its sum, after adjustment, represents  $P_{c_2}^{(10.1)} \times 100$  percent of the total weight resulting from this first step. The estimated proportion

$P_{c_2}^{(10.1)}$ , the Square weight, as well as the weight derived in (3), are used to derive the (1011) and (1001) groups adjusted weight as follows:

$$WE4''_{c_2 i} = WS_{c_2 i} \times P_{c_2}^{(10.1)} \left( \frac{\sum_{i \in n^{(11.1)}} WE4'_{c_2 i}}{\sum_{i \in n^{(10.1)}} WS_{c_2 i}} \right) \quad (5)$$

Consequently, the weight  $WE4'_{c_2 i}$  of both (1111) and (1101) groups must be decreased by the same proportion to respect the overall sum of weight. Equation (6) consists of the final C1-C4 Full file weight prior to the poststratification adjustment. The poststratification is implemented using the same approach as for the Square and Full file weights.

$$WE4''_{c_2 i} = \begin{cases} WS_{c_2 i} \times \frac{\sum_{i \in n^{(11.1)}} WE4'_{c_2 i}}{\sum_{i \in n^{(1.1)}} WS_{c_2 i}} & i \in n^{(10.1)} \\ WE4'_{c_2 i} \times [1 - P_{c_2}^{(10.1)}] & i \in n^{(11.1)} \end{cases} \quad (6)$$

### 3.3.4 Comparison of the methods

Empirical comparisons were made between both methods examined in this paper. First, estimates of mean square errors (MSE) were computed to evaluate the overall quality of the methods. Estimates of biases were defined as differences between C1-C4 Full file estimates and Square file estimates (considered here as the reference). All computations were done using 100 variables from Cycle 1 since Cycle 1 variables were available for all panel members in the Square file. The proposed approach showed a smaller MSE for 53% of the variables. When focusing on the bias, both methods performed equally, which was expected since method #1 adjustments rely only on Cycle 1 variables, and therefore is expected to perform well for these variables.

In terms of assessing the performance of the proposed method for Cycle 4 variables, estimates from the C1-C4 Full file were compared to the estimates from the Cycle 4 Full file, according to the two methods examined. Since the C1-C4 Full file has a significant increase in sample size compared to the Full file, it is not appropriate to make comparisons based on the MSE. However, biases could be compared. Based on 100 Cycle 4 variables, the proposed method showed an average absolute relative bias of 0.9%, compared with 1.5% for results based on the weights computed according to method #1.

## 4. CONCLUSION

As desired while developing the weighting strategy, the proposed method takes advantage of the most recent information available for the different subgroups of panel member part of the C1-C4 Full file. Another important issue was to be consistent with the existing Full file, and this was achieved as the results from section 3.3.4 showed. Empirical comparisons showed a small, but considerable gain with the proposed method, and it is believed that this gain will increase as cycles go by.

## ACKNOWLEDGEMENTS

The authors are grateful to Mylène Lavigne, David Haziza and Sandra Tolusso for their valuable comments that helped improve the quality of the paper. Thanks are also due to Johane Dufour for her useful discussions and comments during the realisation of the project.

## REFERENCES

- Gelman, A. and Carlin, J.B. (2002). Poststratification and Weighting Adjustments, pp. 289-302, in *Survey Nonresponse*: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A (Eds.), John Wiley & Sons, New York.
- Tambay, J.L. and Catlin, G. (1995). Sample Design of the National Population Health Survey. *Health Reports*, 7, 1, 29-38.