Edit and Imputation Processing for Ethnocultural Variables: The Experience of the 2011 Canadian National Household Survey

Sean Crowe, Chunxiao (William) Liu
Social Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

Abstract
The ethnocultural (EC) questions on the 2011 Canadian National Household Survey (NHS), which complements the Canadian Census, measured ethnic and cultural characteristics of the Canadian population. These characteristics are very closely related. For prior Canadian Censuses, the five topics of Immigration and Citizenship, Place of Birth of Parents, Aboriginal, Ethnic Origin and Population Group were processed sequentially through edit and imputation (E&I). In addition to the less-than-optimal efficiency of separate processing, an unacceptable quantity of outlier combinations was present in the imputed data, necessitating a great number of manual post-E&I fixes. For the 2011 NHS, the five separate EC topics were combined into one unified topic with the goal of simplifying the E&I processing, improving the internal coherence of the imputed data and reducing manual intervention after imputation. This paper describes the challenges that were faced and the solutions developed in order to accomplish this task.

Key Words: Canadian Census, Canadian National Household Survey, Ethnocultural variables, CANCEIS, nearest neighbour donor imputation

This paper describes the development of the new edit and imputation (E&I) methodologies for processing the 2011 National Household Survey (NHS) variables and the experience of implementing these. In section 1, some background knowledge is introduced. In section 2, the Ethnocultural E&I situation before 2011 is described. In section 3, the challenges of redesigning and implementing the new methodologies are presented. In section 4, the new strategies and their implementation are detailed. The paper ends with a conclusion in section 5, providing some reflections on the 2011 experience, options considered for 2016 and an overall assessment of the redesign project.

1. Background

1.1 The Canadian Census and the National Household Survey
The Canadian Census program conducts two surveys every five years. The Canadian Census itself consists of a 100% census of all Canadian households and involves questions about demography and language. A supplemental sample survey, named the National Household Survey for 2011, covers several additional topics, such as Labour, Education, Income and Ethnocultural questions.

Prior to 2011, the supplemental sample survey had been a mandatory 1 in 5 sample survey of all Canadian private households and non-institutional collective residences. The

2011 NHS was a voluntary survey with 30% of private households sampled. This came to about 4.5 million households representing approximately 9.6 million Canadians.

1.2 The CANadian Census Edit & Imputation System

The Canadian Census and the NHS have their E&I processing performed by the Canadian Census Edit & Imputation System (CANCEIS), which was developed in the 1990's by Mike Bankier of Statistics Canada (Bankier, 2000). CANCEIS can perform both nearest-neighbour donor and deterministic imputation, as well as derive new variables to use in subsequent modules or in subsequent analysis. Users can fine-tune its functioning through a suite of parameters.

Over the years, the CANCEIS development team at Statistics Canada have continually added features to fulfill the growing needs of users. This was particularly important given the additional functionality required for the revamped E&I processing procedures of the Ethnocultural variables for 2011.

1.3 The Ethnocultural (EC) Variables

There were ten Ethnocultural questions on the 2011 NHS questionnaire. They represented 28 variables processed through E&I.

Prior to 2011, these variables were grouped into five processing subtopics:

1. Immigration and Citizenship questions
    - Place of Birth question
        - Where was this person born?
            - ✓ 13 in-Canada province/territory checkboxes (in-Canada summary variable: POB_IN)
            - ✓ One outside-Canada country write-in (outside-Canada variable: POB_OUT)
        - Overall final summary variable: POB
    - Citizenship question (skipped if person lives on an Indian reserve)
        - Of what country is this person a citizen? (indicate more than one citizenship, if applicable)
            - ✓ "Canada by birth" and "Canada by naturalization" checkboxes (in-Canada summary variable: CIT_IN)
            - ✓ Up to two write-ins for countries outside Canada (outside-Canada variables: CIT_OUT1 and CIT_OUT2)
    - Immigration Status question
        - Is this person now, or has this person ever been, a landed immigrant?
        - "No" and "Yes" checkboxes
        - Overall final summary variable: IMSTAT
    - Year of Immigration question (skipped if Immigration Status is "No")
        - In what year did this person first become a landed immigrant?
            - ✓ Year write-in variable: YRIM
2. Place of Birth of Parents question
    - Place of Birth of Father – part (a) of question
        - Born in Canada checkbox (inside-Canada variable: POBF_IN)
        - Born outside Canada country write-in (outside-Canada variable: POBF_OUT)
        - Overall final summary variable: POBF

- Place of Birth of Mother – part (b) of question
  - Born in Canada checkbox (inside-Canada variable: POBM_IN)
  - Born outside Canada country write-in (outside-Canada POBM_OUT)
  - Overall final summary variable: POBM

3. Aboriginal questions
   - Aboriginal Self-Response question
     - Is this person an Aboriginal, that is, First Nations (North American Indian), Métis or Inuk (Inuit)? (if "yes" then indicate as many as applicable)
     - Four checkboxes:
       - ✓ No, not an aboriginal person
       - ✓ Yes, First Nations (North American Indian)
       - ✓ Yes, Métis
       - ✓ Yes, Inuk (Inuit)
     - Overall final summary variable: ASR
   - Registered Indian question
     - Is this person a Status Indian (Registered or Treated Indian as defined by the *Indian Act* of Canada)?
     - "No" and "Yes" checkboxes
     - Overall final summary variable: RGIND
   - Indian Band/First Nation question
     - Is this person a member of a First Nation/Indian band?
     - "No" and "Yes" checkboxes (checkbox summary variable: BFN_CB)
     - Specify name of First Nation/Indian band write-in (skipped if BFN_CB = "No", variable: BFN_WI)
     - Overall final summary variable: BFN
4. Ethnic Origin question
   - What were the ethnic or cultural origins of this person's ancestors?
   - One to six write-in ethnic origins (variables: ETH1 – ETH6)
5. Population Group (Visible Minority) question (skipped if ASR is "yes")
   - Is this person: (mark more than one or specify, if applicable)
   - 11 check boxes of specific population groups
   - One write-in for "other" population group

Historically, most of the cases requiring imputation have been due to item nonresponse. There are also noninformative invalid (nonsensical) responses that are treated the same as missing information. Additionally, there are informative invalid responses (pseudocodes); for example, pseudoregion responses such as "South America" for Place of Birth or country of Citizenship. These pseudocodes need to be resolved to a value in the appropriate subset of valid responses.

There are not many cases of outright inconsistencies between the values of EC variables that definitely require imputation. Only a few such "impossible" cases arise within the Immigration and Citizenship subtopic, such as the inconsistency of someone being both a Canadian citizen by birth and an immigrant.

However, there are several outlier relationships between EC variables, such as the case of someone identifying themselves as both Aboriginal and also an immigrant. Although

such outliers are theoretically possible, they should be extremely rare. The EC variables are considered to be very closely related. Therefore, it has always been crucial to maintain a certain overall coherence, meaning reasonable counts for all outlier relationships between the EC variables.

2. Ethnocultural E&I Processing Prior to 2011

2.1 Processing Strategy

Before 2011, the five EC subtopics were processed consecutively, i.e., the variables for one subtopic were finalized before starting E&I on the next subtopic. This meant that multiple donors were often used to impute a single failed record, adversely affecting the coherence between the EC variables. In fact, multiple donors could be used for the same failed record within a subtopic, e.g., for the three variables of the Aboriginal subtopic, three different donors could be used to impute for one record.

Coherence between EC variables was managed by looking backward to variables finalized in preceding subtopics and forward to the raw responses of subsequent subtopics. This produced a certain measure of coherence between the final values of the EC variables.

A very important source of ambiguity that is thought to cause high numbers of systematic response errors involves the term “Indian”. The most common type of Aboriginal person is the “North American Indian”. However, this concept can be confused with people being born in the country of India or having an ethnic origin of (east) Indian. Thus, many cases of “false positive” Aboriginal identifications have to be corrected in E&I.

However, decisions made to identify and correct this and other systematic errors were not always applied consistently over all of the subtopics. In addition, the deterministic decision logic tables used to carry out these procedures were sometimes very complex. This led to many manual fixes being required after E&I, which affected tens of thousands of records.

2.2 Commonalities of Approaches Among Subtopics in EC Imputation Prior to 2011

Before 2011, the E&I methodologies for the five EC subtopics were developed by separate teams of methodologists and subject matter specialists. Therefore, the strategies developed for different subtopics were not always totally consistent. However, there were many commonalities in approaches, if not in detailed execution, that will be presented in this section.

2.2.1 Stratification

The input data for CANCEIS deterministic and donor modules can be subdivided into strata which are processed separately. For example, stratification is often used to group similar records together for donor imputation. This was not done for any of the EC subtopics since it was much more important to match on the EC variables with responses, as well as the previously finalized mother tongue (MT) variable. In general, for example, a donor matching very well on the EC variables and MT from the other side of the country was preferable to a donor matching more poorly from across the street.

2.2.2 Use of the Census Family (CF)

Census families are formed prior to Ethnocultural E&I in the Demography topic. This concept was crucial to the pre-2011 strategy of imputing variables in all five EC subtopics. There are four types of CFs used in NHS E&I processing:

- Single person,
- Couple without children,
- Single parent with children, and
- Two parents with children.

It should be noted that a single person is technically regarded as a non-Census family person but is considered a CF of one person for processing purposes.

The first two types of CFs have one generation while the last two have two generations. This leads to the important definitions of CF adult and CF child. The first two CF types involve only CF adults. In the last two types, the parents are CF adults while the children are CF children. No age assumptions are attached to the CF adult/child concepts. For example, a 70-year-old person can be a CF child if they live with their parent(s). On the other hand, foster children can be single persons within a CF, which would make them CF adults under our definition.

2.2.3 Imputation Strategy Using CF Adults and Children

Prior to 2011, there were commonalities, among the five EC subtopics with respect to how the concepts of CF adults and children were used in imputation.

The first observation behind this common strategy is that EC characteristics tend to be much more correlated within CFs than between general members of the population. On the other hand, reported EC variables and highly correlated non-EC, like mother tongue, can be used to find a donor with similar EC characteristics from within the general population. The question then becomes under what circumstances should information for donor imputation come from:

a) Other members of the same Census Family? or
b) The general population?

Clearly this is not a question for single CF persons. For CF couples, the two members of the couple may not resemble each other on EC characteristics, although they generally tend to be much more similar than two unrelated members of the general population. Therefore, the decision was made for all five EC subtopics to use nearest-neighbour CF adults from the general population as donors for other CF adults.

On the other hand, the EC characteristics of CF children tend to be much more similar to those of their CF sibling(s) and parent(s). Therefore, the decision was made to take EC information from other members of the same CF to impute for CF children.

In all cases, the imputation for CF adults was completed first within a subtopic so that a CF child would always have at least one other CF member (a parent) with complete EC information to provide.

2.2.4 All Failed Records Must be Imputed in Every Donor Module

In previous Censuses, it was considered compulsory for 100% of the failed records to be imputed successfully in a CANCEIS donor module. Indeed, in the past CANCEIS would

produce a fatal error if any records could not find a donor for imputation. This had two great drawbacks:

1) Some failed records would be imputed with a very dissimilar donor. In many of these cases, a deterministic resolution would have been much better.
2) The number of related variables that could be imputed in one donor module was limited by the necessity to find a single donor for each failed record with properties suitable for imputing all of these variables at one time. Thus, donor modules tended to impute only one, or at most a few, variable(s) at a time.

3. Challenges for 2011 Redesign

3.1 Processing Changes

We first decided that all 28 EC variables processed through E&I would be treated together under one unified topic for 2011.

This involved two major strategies which improved the final coherence among EC variables:

1) All EC responses would be examined together once to determine when deterministic imputation was warranted to resolve inconsistencies between EC variables and outlier situations arising from probable response error. However, where feasible, using donor imputation instead to resolve inconsistencies would be preferred.
2) A single donor would be used to resolve all the problems with the EC variables for a failed record whenever possible.

These changes would greatly simplify the EC E&I processing. A feasibility study was conducted to determine if, and to what degree, the above strategies could be incorporated for 2011.

Implementing the unified deterministic rules of point (1) would require the coordination of the subject matter experts and methodologists to come to agreements on what deterministic actions to take in the various cases. An example of a problem identified as advantageous to solve through donor imputation rather than deterministically, as in the past, was when the respondent gave both in-Canada and outside-Canada responses for the place of birth of a parent.

With some reasonable simplifying assumptions, the study also yielded simulated results indicating that we could expect over 90% of the failed EC records to be successfully imputed using one donor. Again, the coordination of the subject matter experts and methodology would be required to decide upon the detailed donor imputation strategy and parameters.

3.2 Technical Challenges

3.2.1 CANCEIS Improvements

The revised E&I procedure for imputing the EC variables necessitated several improvements in CANCEIS. Since we would be performing imputation for all 28 EC processing variables at once, CANCEIS would be handling much larger data files than it had in previous Censuses. The software would need to have the capability to access and

manipulate this large volume of data with the speed required to complete this processing within the NHS E&I processing schedule.

In addition, there would be many more different kinds of failed records depending on which of the 28 EC variables had problems to resolve and on non-EC variables such as geography. For example, it had always been a requirement that if Aboriginal variables were imputed by donor for someone living on an Indian reserve then the donor had to come from that same Indian reserve. This geography restriction for the donor did not apply off Indian reserves. Therefore, flexibility was required in assigning imputation parameters such as the relative importance of matching variables used in the nearest-neighbour search (imputation weights). It was not the "one size fits all" situation that may have worked when imputing the variables from the old five subtopics separately.

Finally, the capabilities of specifying a minimum standard for donors and of allowing a portion of the failed records not to be imputed in a donor module would be required. This meant that we would have the freedom to look for an "ideal" donor for the majority of failed records and then deal with the remaining few unimputed records afterward.

3.2.2 Voluntary NHS

There is always an inherent danger of nonresponse bias in voluntary surveys[1]. Thus, the 2011 NHS was designed as a two phase survey where a sample of nonresponders from the first phase was chosen for nonresponse follow-up. In the end, the sampling weights for total nonresponders after the second phase were transferred to responders (Verret, 2013). This left only item nonresponders and inconsistent records for the regular E&I processing.

Although this preliminary treatment ameliorated the nonresponse bias problem, we still anticipated some data patterns and distributions not seen in previous supplemental sample surveys. This could include a significant increase in some types of outliers.

4. Implementation of New E&I Strategies

The new E&I process for the EC variables was divided into two parts:

1) Pre-donor deterministic processing
   - Initial clean up of load data
   - Deterministic imputation where warranted
   - Preparation of imputable and matching variables for donor imputation
   - Modules: ECD1 → ECD2 → ECD3 → ECD4 → ECD5
2) Donor imputation processing
   - CF adults imputed with another CF adult as donor
   - CF children imputed with another member of the same CF as donor
   - Modules: ECI1 → ECD6 → ECI2 → ECD7 → ECI3

Note the naming terminology for the EC modules. The first two characters identify the EC topic. The meaning of the third character is as follows:

"D": Deterministic module
"I": donor Imputation module

[1] The total nonresponse rate for the 2011 NHS turned out to be over 30% compared to about 7% for the 2006 supplemental sample survey (both unweighted).

4.1 Pre-donor Deterministic Processing

The overriding consideration in deciding whether deterministic imputation was warranted was the requirement that a specific systematic error could be reasonably inferred with the information available at that point. Without this strong evidence, data-driven donor imputation was much preferred. For this purpose, we used a systematic hierarchical method:

ECD1:

- Read in load data and create versions of EC variables to be used in E&I.
- E.g., Convert the 13 Canadian province/territory checkbox responses for Place of Birth into the summary variable POB_IN.

ECD2:

- Consider responses within each EC question and perform deterministic imputation if warranted to resolve problems.
- E.g., If the respondent provided "Ontario" as the Place of Birth write-in then we would impute POB_In = Ontario and POB_Out = "valid blank". This action is warranted regardless of the responses to the other EC questions as it is simply an information transfer.

ECD3:

- Consider responses within each EC subtopic and perform deterministic imputation if warranted to resolve between-question problems within the subtopic. The subtopics Immigration and Citizenship, and Aboriginal, were comprised of several questions.
- E.g., Within the Immigration and Citizenship subtopic, if the respondent identifies themselves as an immigrant but also a citizen of Canada by birth who was born in Canada, this was considered sufficient evidence to deterministically change their status to non-immigrant.

ECD4:

- Consider all responses within the EC topic and perform deterministic imputation if warranted to resolve between-subtopic problems.
- E.g., Assume that the respondent identifies themselves as a "North American Indian" within the Aboriginal topic but also reports being born in the country of India, having east Indian but not North American Indian ethnic origin(s), and does not live on an Indian reserve. This was sufficient evidence to identify the "North American Indian" response as a false positive and remove it.

ECD5:

- Monitor the data patterns coming out of ECD4 to ensure that they are expected. The subsequent donor modules would not be able to process unexpected data patterns properly. "Incorrect" data patterns found could be the result of errors in the ECD1-4 modules or of an incomplete understanding of all of the patterns that should be handled by the donor modules.
- Prepare matching variables to use for EC donor imputation.
- E.g., Two immigrant CF parents are missing their Year of Immigration response. They have two children. The older child is an immigrant born in 1984. The younger child is a citizen of Canada by birth who was born in 1991. Therefore, it is reasonable to assume that the parents immigrated to Canada between 1984 and

1991. This restriction on YRIM would then be enforced in the donor modules for CF adults.

We now arrive at the key point in the revised strategy where the great majority of failed EC records would use one nearest-neighbour donor to resolve all remaining EC problems at once.

4.2 Donor Strategy: CF Adults

The donor imputation of EC adults utilized a "two pass" approach. This was made possible by the new CANCEIS functionality to specify a minimum standard for a donor and accept that some failed records may not be imputed in a donor module (first pass).

The first pass consisted of the donor module ECI1.

4.2.1 First Pass

ECI1:

- An essentially "ideal" donor was sought for the failed CF adult. The CF adult donor would have to match exactly, or nearly exactly, on all of the most important EC variables, where responses existed, and the most important non-EC matching variables, e.g., mother tongue and geography. A close match was also required for the remaining matching variables. As previously mentioned, these matching requirements could vary depending on the properties of the failed record.
- E.g., For a failed CF adult living on an Indian reserve and requiring imputation for one or more Aboriginal variables, the donor had to come from the same Indian reserve.
- If an "ideal" donor could not be found in ECI1 then the record would not be imputed in this module. It would need to be imputed in the second pass.

4.2.2 Second Pass

The second pass consisted of the deterministic module ECD6 followed by the second and final donor module for CF adults, ECI2. Only a very small proportion of failed CF adults (well under 10%) were not successfully imputed in ECI1 and required the second pass.

ECD6:

- In this module, the required deterministic imputation was carried out to ensure that either all problems in the failed record were resolved or that an acceptable donor would be found in ECI2 for the remaining problems.
- The deterministic imputation of ECD6 could be required because:
  - No donor could be found to resolve all of the problems of the failed CF adult record properly, or
  - The only donors found were be so dissimilar to the failed CF adult record that deterministic imputation based upon subject matter expertise was superior to donor imputation.
- E.g., Assume a failed CF adult lived on an Indian reserve and required imputation for Indian band/First Nation (BFN). Since the imputation was not done in ECI1, no suitable donor on the same reserve could be found. Therefore, BFN was imputed deterministically in ECD6. A missing BFN would be imputed using the most common BFN in that Indian reserve.

- Deterministic imputation was also performed in ECD6 to resolve all remaining pseudoregions since finding an ECI2 donor to provide a resolving country could not be counted upon.

ECI2:

- In this module, the best (most similar) donor was found to clean up the remaining CF adult problems. There was no minimum donor standard in this module so that all remaining failed CF adult records would be imputed. The deterministic imputation carried out in ECD6 ensured that an acceptable donor, if not always a very similar one, would be found in ECI2.

After the second pass, all EC variables would have been imputed and finalized for the CF adults.

4.3 Donor Strategy: CF Children

4.3.1 General Considerations

With all of the CF adults already imputed, every CF child would be assured of at least one donor (a parent) within the CF at this point. The donor imputation was done in ECI3. Preference was given to using siblings of similar age as donor, if they existed, and given all else being more or less equal.

However, it could occur that no donor in the same CF would be acceptable for some CF children. For instance, if previous deterministic imputation had decided that a CF child was born outside of Canada (but the specific country still had to be determined), and all of the other CF members were born in Canada, then no suitable CF donor would exist to impute for POB.

To deal with these situations, the deterministic module ECD7 was run first. This deterministic module was very similar in function to ECD6 in that it performed the necessary deterministic imputation to ensure that either all problems in the failed CF child record were resolved or that an acceptable donor would be found in the subsequent donor module ECI3 for the remaining problems. This module also derived special "proxy" variables used to impute values to the failed CF child from a donor within the CF.

4.3.2 Proxy variables

When using a sibling as a donor to impute for a CF child, it makes sense to impute the values from the sibling directly. However, if one of two parents is used as the CF donor to impute for the CF child then there are some CF variables where it is not appropriate to take only the values from the one parent. This is the case with the Aboriginal, Ethnic Origin and Population Group variables.

For the Aboriginal subtopic, the CF child should have an Aboriginal characteristic imputed if either one or both parents have that characteristic. For example, if the Registered Indian variable is missing from the CF child, and a sibling donor is not available, then either parent being a Registered Indian is sufficient for the child to be imputed as one; therefore RGIND_PROXY = Yes would be assigned to both parents.

For the Ethnic Origin and Population Group variables, a CF child would normally have the characteristics of both parents. For example, if one parent has the one ethnic origin of

Russian and the other parent has the one ethnic origin of French, and a sibling donor is not available, then the CF child should have both Russian and French imputed as ethnic origins, regardless of which parent is chosen as the donor in ECI3. Hence, ETH1_PROXY = Russian and ETH2_PROXY = French would be assigned to both parents.

5. Conclusions

5.1 Example of Volume of Data Processed for 2011 EC E&I

Here are some counts of the CF adult records processed through the 2011 EC donor modules ECI1 and ECI2:

- 4,629,238 CF adults were edited
- 634, 601 records failed the editing in ECI1
- 600, 092 failed records were successfully imputed in ECI1
- 34, 509 failed records could not be imputed in ECI1
- 6, 609 failed records were completely resolved deterministically in ECD6
- 27, 900 records were successfully imputed in ECI2

5.2 Some Reflections on the 2011 Experience

Direct comparisons to the results of previous Censuses were not appropriate due to the great change in collection methodology of the supplemental sample survey from mandatory response to voluntary. However, the redesign of the edit and imputation procedure for the ethnocultural variables for the 2011 National Household Survey can still be deemed a success:

- The E&I procedure was greatly simplified, which is crucial for future maintenance of the process.

- There were fewer coherence problems among the EC variables after E&I as demonstrated by the subsequent simplified certification of the imputed data, as well as the fewer manual fixes required compared to previous Censuses. This is strong validation of the "one donor" approach.

- The capabilities of the CANCEIS software were greatly enhanced by the impetus of this project. The ability to process much larger data sets in a timely manner was needed for some other 2011 NHS topics, for example, for the E&I processing of the Income topic. Similarly, the flexibility of assigning different imputation parameters to different types of failed records was used in the E&I processing of several 2011 NHS topics.

5.3 Options Considered for the 2016 NHS

There were still a few manual fixes required in 2011 as a result of unacceptably high counts for some outlier combinations of EC variables after E&I was completed. For 2016, we need to ensure that the outlier counts are closely monitored from the start of E&I processing, and that appropriate actions are built into the E&I modules when the counts are unacceptable.

In addition, although the two people who make up a CF couple may be quite different in their EC characteristics, they are generally much more similar than two random members of the CF adult population. Therefore an effort should be made to impute CF couples

using other CF couples as donors. In other words, CF adults would be split into two strata for donor imputation:

1) CF single persons, and
2) CF couples, where the unit of imputation would have two subunits.

This would much better reflect the inter-couple EC characteristics inherent in the population.

Finally, we should investigate whether the 2011 EC E&I strategy can be generalized to other topics for 2016.

5.4 Overall assessment of the Redesign Project

Although the development of the EC E&I took significantly longer and required more resources for 2011 than in previous Censuses, the ongoing benefits are many and very important as summarized above. In particular, the simplification and improved quality of the E&I process will serve EC processing well for future maintenance. In addition, the insights gained will continue to yield benefits well beyond the Ethnocultural topic. This project was an important investment in the E&I processing infrastructure or, to use the colloquial term: "Short-term pain for long-term gain!"

Acknowledgements

The authors would like to thank the ethnocultural subject matter specialists Jarod Dobson and Laetitia Martin, and our colleague methodologists Lyne Guertin and Marcel Bureau for their great assistance in the creation of this paper.

References

Bankier, M. (2000), "Imputing Numeric and Qualitative Variables Simultaneously", Social Survey Methods Division Report, Statistics Canada, Dated February 21, 2000.

Crowe, S. and Liu, W. (2013), "2011 NHS Ethnocultural Topic: Donor Imputation Narrative", Statistic Canada internal document.

Verret, F. (2013), "The Estimation Methodology of the 2011 National Household Survey", Proceedings of the 2013 Joint Statistical Meetings.