Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Overcoverage in the 2011 Canadian Census

Abel Dasylva, Robert-Charles Titus and Christian Thibault[1]

Abstract

The Census Overcoverage Study (COS) is a critical post-census coverage measurement study. Its main objective is to produce estimates of the number of people erroneously enumerated, by province and territory, study the characteristics of individuals counted multiple times and identify possible reasons for the errors. The COS is based on the sampling and clerical review of groups of connected records that are built by linking the census response database to an administrative frame, and to itself. In this paper we describe the new 2011 COS methodology. This methodology has incorporated numerous improvements including a greater use of probabilistic record-linkage, the estimation of linking parameters with an Expectation-Maximization (E-M) algorithm, and the efficient use of household information to detect more overcoverage cases.

Key Words: census, record-linkage, overcoverage, Expectation-Maximization (E-M), administrative sources, record group.

1. The Census Overcoverage Study

In the census of population, overcoverage occurs whenever the same person is enumerated multiple times. In 2011, important changes were introduced to the census process, which increased the likelihood of overcoverage. Those changes included optional long forms for the National Household Survey, a wave methodology to encourage online responses and an increased proportion of mail-out.

As is the case with undercoverage, overcoverage has a major impact on the accuracy of census counts. For this reason, it is measured by the Census Overcoverage Study (COS) based on multiple enumerations within the in-scope population. In 2011, the in-scope population included all Canadian citizens and landed immigrants having either a usual place of residence in Canada or living abroad on a military base, or diplomatic mission, as well as those at sea or in port, aboard a merchant vessel registered in Canada, on Census day. The in-scope population also included non-permanent residents and family members living with them, if their usual place of residence was in Canada, and they were either claiming refugee status or holding a valid permit to study or work, for a period covering Census day. The COS estimate is an important input for the net undercoverage, which is defined as the difference between the undercoverage and the overcoverage. It is also an important input for Statistics Canada Population Estimates Program.

The 2011 COS is designed as a sample survey, where overcoverage is estimated by a probability sample drawn from a frame of *potential overcoverage cases*, which are groups of census records connected through record-linkage even though the records may not represent any actual overcoverage. Therefore the COS includes all the steps found in typical sample surveys, from the construction of the survey frame to the estimation.

However, the COS is also an unusual survey because its frame is built through probabilistic record-linkage (Fellegi and Sunter, 1969). In fact, the COS frame is the union of three overlapping frames; the *Step 1 frame*, the *Step 2 frame* and the *Extension frame*. The Step 1 frame is built by linking the census response database to an administrative file and forming the record-groups that are connected by the resulting links. The Step 2 frame is built by linking the census records that are not linked in Step 1, to the entire census response database, and building

[1] Abel Dasylva, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A0T6, Canada (abel.dasylva@statcan.gc.ca); Robert-Charles Titus, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A0T6, Canada (robert-charles.titus@statcan.gc.ca); Christian Thibault, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A0T6, Canada (christian.thibault@statcan.gc.ca)

groups of connected records with the resulting links. Finally, the Extension frame is built by exploiting a household identifier that is available on the census response database. Two households are linked if they contain census records that have been linked in Step 1 and Step 2. For such a pair of linked households, additional links are created between their census records. The Extension frame is comprised of all groups of records that are connected by such links.

Another important difference with a typical sample survey is in the data collection phase, where a sample of record-pairs is reviewed manually and respondents are not involved. During this clerical review, each link in a potential overcoverage case is verified. The results of these verifications are then processed to produce *verified overcoverage cases*, which are groups of records connected by links with verified overcoverage.

The 2011 COS has incorporated many improvements to the 2006 methodology, including an Expectation-Maximization (E-M) algorithm to determine the linkage weights as proposed by Winkler (1988) and Jaro (1989), provincial/territorial weight thresholds and the detection of additional overcoverage based on linked household-pairs.

The following sections are organized as follows. Section 2 describes the input files. Section 3 describes the sampling frame. Section 4 describes the sampling design. Section 5 describes the processing and estimation. Section 6 presents the results.

2. Input files

The input files were the Census Response Database (RDB) and an Administrative Frame (AF) built from many sources.

2.1 Census RDB

The RDB includes responses from people living in private or collective dwellings. In 2011, it had 32 million records and included the following information:

- Names: given names and surnames as two separate variables
- Demography: birthdate and sex
- Geography: province/territory, postal code, census division, collection unit and civic address

Given names and surnames were parsed into components and standardized.

2.2 Administrative frame

The AF is built to provide the largest coverage of the target population. In 2011, it had 47 million records and included the same name and demographic information as the RDB. However the geographical information was limited to the province/territory and the postal code of the mailing address.

The AF included the records (hereafter called *administrative records*) from the following sources:

- T1 Personal Master tax Files (T1PMF) from 2005 to 2009. They were provided by Canada Revenue Agency (CRA) and represented 58.2% of the administrative records.
- Canadian Child Tax Benefits (CCTB) files up to July 2011. They were also provided by CRA and represented 15% of the administrative records.
- Birth records from Vital Statistics files from 1974 to 2008. They were provided by Statistics Canada Health Statistics Division and represented 12.2% of the administrative records.
- Immigrants and Non-Permanent Residents files up to September 2011. They were provided by Citizenship and Immigration Canada (CIC) and represented 14.4% of the administrative records.
- Territorial Health Care Files (HCF) up to July 2011. They were provided by the territories and represented 0.2% of the administrative records.

Note that the addition of the CCTB in 2011 greatly increased the AF coverage, as compared to 2006.

Each administrative source was unduplicated separately. However, the AF had duplicates across the different sources, because no unique identifier covered all the sources. Consequently the AF had more records than the RDB. To address this issue, the links between a given census record and the AF were prioritized as follows. For an adult (i.e. over 18 on Census day) from the province (i.e. not leaving in a territory), links to T1PMF records had the

highest precedence, followed by links to CCTB records and then links to CIC records. In that scheme, a link with a given precedence was ignored in favour of a link with a higher precedence. Different priorities were used for a child in a province; links to CCTB records were given the highest priority followed by links to birth records. Finally, all territorial census records were only linked to records from the territorial HCF.

3. Sampling frames of potential overcoverage cases

The 2011 COS built three sampling frames of potential overcoverage cases, including the Step 1 frame based on the RDB-AF linkage, the Step 2 fame based on the linkage between the residual RDB and the entire RDB, and the Extension frame. Each frame contained groups of census records, which were connected by links.

3.1 Step 1 frame

In this step, the RDB and AF were linked through probabilistic record-linkage with G-LINK; Statistics Canada generalized record-linkage system. That linkage enabled the identification of most potential overcoverage cases, where two or more RDB records were linked to the same administrative record, according to the prioritization scheme described in Section 2.2. It also enabled the identification of pseudo-duplicates. These are census records that happen to agree on many linkage variables but represent different individuals.

3.1.1 Blocking criteria

Blocking was required given the sizes of the input files, including 32 and 47 million for respectively the RDB and AF. It was based on the Soundex code of the names, birthdate components including transpositions of month and day of birth and the postal code.

3.1.2 Rules and their outcomes weights

Six rules were used for the detailed comparison of records within pairs. They included the following:

- *First two surnames*: a matrix rule with partial agreements (including exact and typo in that order) and transpositions
- *First two given names*: a matrix rule with partial agreements (including exact, typo and alternate name in that order) and transpositions
- *Day and month of birth*: a matrix rule with exact comparisons and transpositions
- *Year of birth*: exact comparison
- *Postal code*: exact comparison
- *Sex*: exact comparison

The rules outcomes weights were estimated with an E-M algorithm based on assumptions of conditional independence, as proposed by Winkler (1988) and Jaro (1989). The algorithm incorporated three enhancements. Firstly, the algorithm directly estimated the u-distribution from a sample of random pairs, outside the E-M algorithm. This means that only the m-probabilities and the mixing proportion were estimated iteratively by the E-M algorithm. Secondly, the estimated u-distribution accounted for interactions among the unmatched pairs satisfying the blocking criteria. Thirdly, the E-M algorithm accounted for missing linkage variables as explained in the next section.

3.1.3 Missing linkage variables

Linkage variables were missing in some pairs. This missing data may be viewed as a form of item nonresponse, which complicated the estimation of the linkage weights. The E-M was enhanced to account for the missing values under the assumption of Missingness Completely At Random. For the simple rules, which involved the birth year, postal code or sex, that choice lead to a null outcome weight, whenever the corresponding variable was missing in one of the records. The situation was more complex for the matrix rules, where the outcome weight was not always null depending on the specific missingness pattern. That solution was an improvement over previous heuristics that simply assigned the same default value (e.g. null) to each rule with a missing value (Samuels, 2011).

3.1.4 Provincial/territorial names frequency weights

Frequency weights accounted for the relative frequencies of the names. They were added to a pair linkage weight when the outcome was an agreement for the corresponding matrix rule. Those weights were computed according to the provincial/territorial frequency F of the corresponding name, based on the formula $-10\log_2 F$. The province/territory of a pair was determined by the census record. Provincial/territorial frequency weights were crucial because names distributions differ across Canada and especially in Quebec.

3.1.5 Linkage decision based on Provincial/territorial upper thresholds

For each province and territory, a separate upper-weight threshold was computed. It was based on a 1% conditional nonmatch probability given that a pair satisfied the blocking criteria and had its weight above the threshold. The threshold was computed based on the E-M algorithm output. RDB-AF pairs with a linkage weight above their provincial/territorial upper-threshold were linked. Note that the province/territory of a pair was determined by the census record. Finally, redundant links were removed based on the prioritization scheme of Section 2.2.

3.1.6 Record-groups

Mutually exclusive groups of connected census and AF records were formed with the remaining pairs above the upper-threshold. The overwhelming majority of groups were one-to-one groups, i.e. RDB-AF pairs. One-to-one and one-to-many (one RDB record linked to many AF records) cases did not represent overcoverage. The remaining groups represented the potential overcoverage cases of the Step 1frame.

3.2 Step 2 frame

The majority of census records were linked to the AF in Step 1. The remaining unlinked RDB records formed the *residual RDB*. Overcoverage in that part of the RDB was missed in Step 1 possibly because the corresponding person was not covered by the AF. Step 2 detected some of that overcoverage by linking the residual RDB with the entire RDB. The record-linkage methodology in this step was similar to that in Step 1 except for minor changes.

3.2.1 Provincial/territorial names frequency weights

Provincial/territorial frequency weights for names were also used. However national frequencies were used in pairs with records from different provinces or territories.

3.2.2 Linkage decision based on Provincial/territorial lower thresholds

For each province/territory, a separate lower-weight threshold was computed based. It was based on a 1% conditional probability given that a pair had its weight below the threshold given that it was matched. As before, the threshold was computed based on the E-M algorithm output. A single lower threshold was computed for all pairs where the census records came from different provinces or territories. Pairs with a linkage weight above their provincial/territorial lower-threshold were linked.

3.2.3 Record-groups

Mutually exclusive groups of connected records were formed, with the pairs that were above their lower weight threshold. Those groups included an overwhelming majority of pairs. All the groups represented potential overcoverage cases of the Step 2 frame.

3.3 Extension frame

In the past, the double enumeration of entire households has represented an important portion of the overcoverage. To detect a larger part of this overcoverage, additional links were created between census records based on household identifiers present on the RDB. Those links were created in two steps. The first step created a link between two households if they were associated with two census records linked in Step 1 or Step 2. The second step created new links between the other records associated with the linked households by comparing them based on the sex and birthdate. Those new links were then used to create connected groups of census records. These records which represented potential overcoverage cases of the Extension frame.

4. Sampling design

Three independent stratified samples were drawn from the three frames. Those samples were reviewed manually, to verify the occurrence of overcoverage. The clerical review of a potential case was broken into the review of its constituent links. For each such link, the review involved the comparison of the selected records and that of the corresponding households, for better clerical decisions.

The sequential construction of the different frames, and the independent samples gave rise to some overlap as some census records were included in multiple potential cases, from different frames. This issue was addressed at the estimation phase as follows. Firstly, each potential case was included in a larger record-group, called *overlap groups*. Within the same overlap group, records may be connected by links from Step 1, Step 2 or the Extension. Secondly, the sampling weights were adjusted with weight-shares by viewing the union of the three samples as an indirect sample of overlap groups. Lavallée (2002) describes the Generalized Weight Share Method for indirect sampling.

4.1 Strata

In each frame of potential overcoverage cases, the record-groups were stratified by their number of records and the province/territory of each record. The size measure $\hat{P}(M|\gamma)\left[1-\hat{P}(M|\gamma)\right]$ was used to further stratify the Step 2 pairs, where $\hat{P}(M|\gamma)$ is the estimated conditional match probability $\hat{P}(M|\gamma)$ and γ the vector of outcomes observed in the pair. The conditional match probability was computed with the E-M algorithm.

4.2 Sample selection

Within the strata of each frame, the record-groups were first sorted by the sex and birthdate. Then a systematic sample was drawn.

4.3 Sample allocation

In Steps 1 and 2, the allocation was optimized subject to a maximum coefficient of variation for the provincial/territorial estimate of overcoverage. Strata with large groups were designated as take-all. In Step 2, the allocation used the size measure $\hat{P}(M|\gamma)\left[1-\hat{P}(M|\gamma)\right]$ for strata containing pairs.

In the Extension frame, an equal sample size was allocated to each stratum.

5. Processing and estimation

Two sets of estimates were produced including pre-adjustment estimates and final estimates. The final estimates included the overcoverage missed by the COS but found by the Automated Match Study (AMS).

5.1 Processing

The clerical review results were processed into verified overcoverage cases; i.e. groups of census records that were connected through links with verified overcoverage. In a verified overcoverage case, the overcoverage was counted as the number of records minus one. The overcoverage of a sampled potential case was set to the total overcoverage across the included verified cases.

5.2 Pre-adjustment estimates

The overcoverage estimates were based on the Horwitz-Thompson estimator, with some reweighting to account for any overlap.

5.3 Adjustment

The AMS is an evaluation study that is based on households. A small proportion of the overcoverage may be detected by the AMS but missed by the COS. In 2011, for each province and territory, a separate adjustment added the missed overcoverage to the pre-adjustment estimate to produce the final COS estimate.

6. Results

In 2011, the COS estimated the Canada-wide overcoverage at 632,846 with a standard error of 6,675 (Dasylva, 2013). The overcoverage rate was estimated at 1.85%. This was an increase over the 2006 estimate of 1.59% (Statistics Canada, 2006).

Table 6-1 shows the overcoverage by step (Dasylva, 2013). The majority of the overcoverage was detected in Step 1 and Step 2. The Extension detected 5.05% of the overcoverage. As for the adjustment, it only accounted for 1.90% of the final overcoverage estimate.

Table 6-2 shows the distribution of the overcoverage by type, as well as by scenario for the cases involving different households (Dasylva, 2013). The overcoverage between identical households represented the majority of the overcoverage at 51%. The remaining overcoverage (48%) involved different households, with 29% of that coverage representing children in shared custody.

Table 6-1
Overcoverage by steps

Step	***Total***	***%***
Steps 1 & 2	588,856	93.05
Extension	31,939	5.05
Adjustment	12,051	1.90
Total	632,846	100

Table 6-2
Overcoverage by type and by scenario

Type	***%***
Identical, 1 person, far	1
Identical, 1 person, near	4
Identical, multiple people, far	7
Identical, multiple people, near	39
Non-identical, multiple people, 1 in common	28
Non-identical, multiple people, ≥ 2 in common	20
Missing	<1

Scenario of overcoverage among different households	***%***
Child(ren) of parents in separate households	29
Adult with other relatives	17
Student/young adult newly away from home	15
Child(ren) with two relatives/adults	6
Adult entering or leaving married/common-law	5
Adult with other unrelated adults	5
Young adult entering married/common-law from home	4
One household not a private dwelling	3
Other	16
Missing	1

References

Dasylva, A. and Titus R.-C. (2013), "2011 Census Overcoverage Survey (COS) Methodology Report", internal report, Ottawa: Statistics Canada.

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory of Record Linkage", *JASA*, 64, pp. 1183-1210.

Jaro, M. A. (1989), “Advances in record linkage methodology to matching the 1985 census of Tampa, Florida”, *JASA*, 84, pp. 414-420.

Lavallée. P. (2002*), Le Sondage indirect ou la méthode du partage des poids*. Bruxelles: Éditions de l’Université de Bruxelles.

Samuels, C. (2011), ”Using the EM algorithm to estimate the parameters of the Fellegi-Sunter model for data linking”, report number 1352.0.55.120, Australia: Australian Bureau of Statistics.

Statistics Canada (2010). *2006 Census Technical Report: Coverage*. Statistics Canada Catalogue no. 92-567-X. Ottawa, Ontario.

Winkler, W.E. (1988), “Using the EM algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage”, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 667-671.