ESTIMATION OF OVERCOVERAGE IN THE CENSUS OF CANADA USING AN AUTOMATED APPROACH

Claude Julien, Statistics Canada
Ottawa, Ontario, Canada K1A 0T6

KEY WORDS: Coverage evaluation, two-phase design, stratification

1. Introduction

In any survey or census, errors due to the coverage of the target population may occur and thus affect the accuracy of the estimates or counts. It is therefore important to measure these errors and to study the reasons for their presence.

An error due to the omission of a unit of the target population is referred to as undercoverage. Conversely, when a unit is enumerated more than once or a unit not belonging to the target population is enumerated at least once, there is said to be overcoverage.

As part of the 1991 Census of Canada Coverage Error Measurement Program, an Overcoverage study (see Dibbs and Royce 1990) is currently being developed. This study uses several methods to detect and estimate different types of overcoverage:

1. a post-censal survey of households to identify fictitious and out-of-scope persons, and to collect additional addresses where persons may have been double-counted;
2. a survey of usual residents in collective dwellings to identify those who are also enumerated in a private or another collective dwelling;
3. an Automated Match Study to identify overcoverage caused by errors occurring during the census data collection operation.

This paper describes the Automated Match Study (AMS). The next two sections present the objectives of the AMS and how it combines automated and clerical operations to fulfil them. The last two sections describe and evaluate three sample designs that combine both operations to produce an estimate of total within EA overcoverage that is as precise and exact as possible.

2. Objectives of the AMS

The Canadian Census of Population is conducted by dividing the country into approximately 45,000 Enumeration Areas (EA). In general, an EA is an area for which one Census Representative (CR) is responsible. The CR is given a map of the EA and is required to identify and list all the dwellings in the Visitation Record (VR). In the majority of EAs, the CR leaves a census questionnaire at each household to be completed by one of the residents and mailed back on Census Day. Households that do not mail back their questionnaire are followed up by telephone or by personal visit two or three weeks after Census Day.

During the Census data collection operation, some households can be enumerated more than once. For example, a household might mail back its questionnaire a few days late and then complete another during telephone follow-up. Duplication may also occur when a CR unknowingly drops off two questionnaires at the same dwelling. In this case, the residents of the dwelling might complete and return both questionnaires, or complete one and answer the other during subsequent follow-up operations. Within EA overcoverage occurs when the CR fails to detect the duplication.

Although the questionnaires and the VR pass several quality checks, the overcoverage can be left undetected and therefore some persons are present more than once on the census database from which counts are tabulated. The objectives of the AMS are (a) to detect within EA overcoverage as efficiently and effectively as possible, and (b) to estimate total within EA overcoverage as precisely and exactly as possible.

3. Automated and clerical operations

3.1 Description

In the Canadian Census of population, the names and addresses of respondents are not captured. In this context, manually searching all questionnaires in an EA for double-counting is a costly, tedious and error-prone operation. A totally automated approach is not feasible either. The AMS approach combines both strategies. In the first step, a computer program extracts information from the census database and reports pairs of households that are similar enough to possibly include common persons, ie. overcoverage. In the second step, the census questionnaires completed by these households are verified by a clerk who reports the presence or absence of overcoverage.

Within EA overcoverage is more likely to occur among similar households enumerated in the same neighbourhood. In the AMS, similarity and proximity of two households is determined by a specially designed computer program. This program compares the sex and the date of birth of the household members and produces the following statistics: the size of each household, the number of similar persons and the proximity of the households.

These statistics are used to classify a pair of households according to the likelihood that it contains overcoverage. For example, a pair of four-person households with four similar persons is put into a high likelihood class, whereas a pair of four-person households with only one similar person is put into a low likelihood class.

The comparison is done for each pair of households in an EA. The average EA contains 300 households and yields 44,850 comparisons. It is therefore impractical to manually verify each pair. Depending on the class, all or some pairs are selected and printed on a form. The characteristics of all the members are printed side by side for each household. In the verification operation, clerks are assigned to look at the census questionnaire for each household and to indicate on the form which persons are double-counted.

3.2 Feasibility of the AMS

Using 1986 Census data, we evaluated the feasibility of the AMS methodology (see Julien, 1991). In the study, we carried out the automated matching operation in 380 EAs. Two persons from different households were similar when both had the same sex, month of birth and year of birth (the day of birth was not used because it was not available on the database). Two households were considered to be in the same neighbourhood when their household numbers differed by five or less. A household number is given by the Census Representative when canvassing the EA. Table 1 gives the

average number of pairs per EA for each class.

In 40 of the 380 EAs, all pairs in classes 1 to 5 were verified. Table 1 provides the incidence of overcoverage per class (I^c), defined as the ratio of the number of pairs with overcoverage divided by the number of pairs verified. The results reveal that neighbouring households with more than one similar person are almost all cases of overcoverage. In the other classes, the incidence of overcoverage varies between 1 % and 50 %. The fact that the incidence of overcoverage varies substantially among the classes demonstrates the efficiency of the automated operation.

The feasibility study also identified two weaknesses of the AMS methodology; verification in large classes and response errors. Pairs in classes 6 and 7 were not verified because they contain too many pairs of households. The incidence of overcoverage is expected to be very low and very many pairs would have to be verified in order to observe just one case of overcoverage. This problem can be handled by ignoring these classes totally, and tolerating a slight underestimate of total within EA overcoverage, or by verifying a sample of pairs, and obtaining an unbiased yet potentially imprecise estimate.

The AMS methodology relies on the assumption that persons enumerated more than once present similar characteristics in each enumeration, ie. the same sex and date of birth is reported. The study showed that 15 % of the persons enumerated more than once had a different date of birth reported. Consequently, it is expected that a few cases of overcoverage fall in class 6 or 7, instead of falling in classes 1 to 5 where they would be easier to detect.

It is worth noting that the AMS is carried out at the household level. A pair of households that contains overcoverage will present no similarity, and thus fall into class 7, only if all overcovered persons have their date of birth reported erroneously. Fortunately, the chance that such a situation occurs decreases quickly as the number of overcovered persons increases. One potential improvement is to relax the criteria used to determine similar persons at the expense of increasing the number of pairs in classes 1 to 6. Unfortunately, this leads to the first weakness of dealing with larger classes.

4. Three sample designs

The ultimate goal of the AMS is to produce an unbiased estimate of total within EA overcoverage with a specified level of precision by allocating the available resources between the cheap automated matching operation and the expensive clerical verification operation. In order to obtain a reasonable level of precision with a sensible amount of resources we decided to exclude the unlikely pairs falling into classes 6 and 7 from the verification operation. Consequently the target population is not completely covered, but the bias is expected to be small (between 1 % and 5 %).

In this section, we compare three sample designs: a simple random sample design (SRS), a two-phase design using a stratified estimator (TP_STR) and a two-phase design using a ratio estimator (TP_RAT). The notation and formulas employed hereafter are described in the Appendix.

In the SRS, all EAs that are selected for the automated matching operation are also selected for the manual verification operation. The overcoverage observed is simply multiplied by the inverse of the sampling rate to produce an unbiased estimate. Since the verification operation is expensive and time consuming, the number of EAs verified will be small. Given the rareness of overcoverage, the SRS with a small sample is expected to yield a very imprecise estimate.

Since it is much cheaper to process an EA through the matching operation than it is to verify the resulting pairs, the rationale of the two-phase approach is to verify fewer EAs than by the SRS method and to use the extra resources available to match a much larger first phase sample of EAs. The results of the first phase sample are then used as auxiliary information to obtain more precise estimates than the SRS approach.

The TP_STR approach utilizes the results of the first phase sample to distinguish two or more strata of EAs in which the proportion of overcoverage differ greatly. The results are also used to estimate the stratum weights. The second phase sample of EAs is verified to estimate the average number of overcovered persons per EA in each stratum. The estimated weights and averages are combined to produce an unbiased estimate that is expected to be more precise than the SRS estimate. The stratification provides a better use of the resources available for verification, ie. it enables the disproportionate allocation of the second phase sample. For example, the strata consisting of EAs with highly likely pairs would be allocated a relatively bigger share of the second phase sample.

An alternative approach is to consider the population of all within EA pairs of households and to estimate the total within EA overcoverage in each class (Y^c). A difficulty arises because the size of the population in each class, M^c the number of pairs, is unknown. A two-phase design with a ratio estimator offers an attractive solution.

In this approach the results from the first phase sample are used to estimate the M^c. A second phase sample of EAs is verified to estimate the average number of overcovered persons per pair in the c^{th} class. This is done by using the ratio estimator $\hat{Y}^c / \hat{M}^c$. The estimated sizes and averages are combined to produce an estimate that is expected to be more precise than the SRS estimate. However, it is also expected to be biased because of the use of the ratio estimator.

5. Evaluation of the sample designs

In this section, we describe a simulation study in which the three sample designs were compared to evaluate (a) how they perform with such a rare population, (b) the gains of the two-phase approaches and (c) the bias incurred by the TP_RAT approach.

5.1 The population

As mentioned in section 3, 1986 Census data from 380 EAs were processed through the computer matching program. The pairs produced were classified into the seven classes shown in Table 1. A manual verification of pairs in classes 1 to 5 was carried out for 40 of the 380 EAs to determine the incidence of overcoverage in each class(I^c), presented in Table 1.

Using the results of the verification, the presence of overcoverage was simulated for the 340 EAs that were not verified. A random number between 0 and 1, r_{ij}^c, was generated for each j^{th} pair of the i^{th} EA in classes 1 to 5.

When $r_{ij}^{c} < l^{o}$, overcoverage was determined and the number of overcovered persons for that pair, y_{ij}^{c}, was set to the size of the smallest household of the pair; otherwise, no overcoverage was determined and y_{ij}^{c} was set to 0. The 40 EAs that were verified plus the 340 EAs that were simulated made up a population of more than 233,000 persons of which 314 were overcovered. More information on this population is given in Table 2.

5.2 The sample selection

From this population, an initial sample of n' EAs was selected. This sample was used as a first-phase sample for the TP_STR and TP_RAT methods. For the TP_STR method, the number of pairs in each class for each EA (M_i^1, M_i^2, ... , M_i^7), was used to divide the selected sample into 3 strata: stratum 1 was all EAs with at least one pair of households in CLASS 1 ($M_i^1 > 0$), stratum 2 was EAs with no pair in CLASS 1 but at least one pair in CLASS 2 ($M_i^1 = 0$; $M_i^2 > 0$), and stratum 3 was all other EAs. Using the population statistics presented in Table 2, optimal values of the second-phase sampling fractions v_h (see Cochran, 1977, p. 331), were calculated under the assumption that the cost of verification is the same for each stratum and 10 times higher than the cost of the automated matching operation. These sampling fractions were applied to the number of EAs observed in each stratum (n'_1, n'_2 , n'_3) to obtain the second-phase sample size that was selected from each stratum ($n_h = v_h\, n'_h$). For the TP_RAT method, the sum of the n_h gave the second-phase sample size that was selected from the initial sample.

In order to compare designs that are cost-equivalent, combining the automated and clerical operations, the SRS sample size was set to $n_{SRS} = (n' + 10\ n_{TP}) / 11$ rounded to the nearest integer. The SRS sample was selected from the initial sample independently from both second phase samples and thus was equivalent to a random sample taken from the whole population.

5.3 Presentation of the results

The selection method described in section 5.2 was carried out 300 times each for first-phase sample sizes of 80, 120, 160 and 200 EAs. With each simulation, $\hat{Y}_k$ the estimate of the number of overcovered persons and $cv(\hat{Y}_k)$ the estimate of the coefficient of variation were calculated for each sample design (k = 1,2,3).

The results of the simulation are given in Table 3. For each design the average of the 300 $\hat{Y}_k$ and $cv(\hat{Y}_k)$ are provided, as well as the actual coverage rate of the 95 % confidence interval. The latter statistic was calculated by computing, for each simulation, the 95 % confidence interval estimated by each option and counting the proportion of the intervals actually covering the true population value (314). This statistic is expected to be close to .95 for the unbiased SRS and the TP_STR designs. It should also point out any significant bias resulting from the TP_RAT. Furthermore, to show the direction of the bias Table 3 also gives the proportion of confidence intervals that are too low (underestimation) and too high (overestimation).

The 300 estimates of total within EA overcoverage produced by each design averaged close to the population value of 314. The TP_STR estimate averaged closest to the population value with all sample sizes. The TP_STR also yielded the most precise estimate. Its average estimated coefficient of variation (cv) was at least 34 % lower than the cv of the SRS estimate and at least 12 % lower than that of the TP_RAT estimate. However, one has to keep in mind that the TP_STR design was evaluated under optimal conditions, ie. the optimal second-phase sampling rates were known.

The observed coverage rate of the SRS and TP_STR designs are very similar and slightly lower than the expected 95 % coverage rate. This might be caused by the rareness of the population, shown by the highly skewed distribution of Y_i given in Table 2. The coverage rate of the TP_RAT design is much lower, especially with small samples. This indicates that the ratio design tends to underestimate the variance. This design also produced five times more cases of overestimation than the other two designs.

6. Conclusion

The objectives of the AMS are (a) to detect overcoverage occurring within an Enumeration Area as efficiently and effectively as possible, and (b) to estimate total within EA overcoverage as precisely and exactly as possible. The first objective was met by combining an automated matching operation with a clerical verification operation. The former classifies pairs of households according to the likelihood that they contain overcoverage. The latter reports the presence or absence of overcoverage by verifying the census questionnaire completed by the most "suspicious" pairs. This method is very effective in that it isolates most of the overcoverage in a few relatively small classes of pairs of households. However, in order to be efficient the pairs falling in the largest and least likely classes must be ignored, ie. excluded from the verification operation. Consequently the target population is not completely covered, although the bias is expected to be small.

The second objective was met by comparing three sample designs in a simulation study. A two-phase design with a stratified estimator was the best of the three options. In this design, a large first phase sample of EAs is processed through the automated matching operation. Using the results from this operation, the EAs are stratified according to the likelihood that they contain overcoverage. Disproportionate sampling is then applied in the second phase. A bigger share of the second phase sample is allocated to the strata of EAs that are more likely to contain overcoverage. The second phase sample of EAs is verified to estimate the average number of overcovered persons per EA in each stratum. These averages are combined with the estimated stratum weights to yield an unbiased estimate.

The AMS will be carried out sometime between November 1991 and April 1992. To get an idea of the number of EAs to process through the automated matching and clerical verification operations, we assumed that the statistics presented in Table 2 applied to the population of 45,000 EAs and calculated the first and second phase sample sizes required to achieve a specified level of precision. The results are shown in Figure 1. To obtain a coefficient of variation of 10 % we need to match a first phase sample of 788 EAs and verify a second phase sample of 175 of them. A cv of 5 %

requires a first phase sample of 3059 EAs and a second phase sample of 681 EAs.

Currently another simulation is under way to implement the sample design at the province level and to estimate the optimal second-phase sampling rates.

ACKNOWLEGMENT

The author would like to ackowledge Don Royce, Dave Dolson and Ruth Dibbs for their useful comments; and Laurie Reedman for her technical and programming assistance.

REFERENCES

Cochran, W.G., (1977). Sampling Techniques. New York, John Wiley & Sons.

Dibbs, R. and Royce, D. (1990), "Measuring Overcoverage in the 1991 Census of Canada", Proceedings of the Government Statistics Section, American Statistical Association, 24-27.

Julien, C. (1991), "Assessing the Feasibility of an Automated Match Study to Estimate Overcoverage in the Census", Technical Report, Social Survey Methods Division, Statistics Canada.

Kovar, J., Ghangurde, P., Germain, M.-F., Lee, H., and Gray, G. (1985), "Variance Estimation in Sample Surveys", Methodology Branch Working Paper No. BSMD 85-049E, Statistics Canada.

APPENDIX: Notation and formulas for the estimate of total within EA overcoverage and the estimate of variance

Let N denote the size of the EA population, n' the size of the first phase sample, n the size of the second phase sample; let M denote the number of pairs and y represent the number of overcovered persons; let i, c, j and h respectively denote the EA, the class, the pair and the stratum.

$Y = \sum_{i=1}^{N} \sum_{c=1}^{C} \sum_{j=1}^{M_i^c} y_{ij}^c = \sum_{i=1}^{N} Y_i$ is the total within EA overcoverage

Simple random sampling, $\hat{Y}_1 = N\bar{y} = N\dfrac{\sum_{i=1}^{n} Y_i}{n}$; $v(\hat{Y}_1) = N^2\left(\dfrac{1}{n} - \dfrac{1}{N}\right) s^2(y)$, where $s^2(y) = \dfrac{\sum_{i=1}^{n}(Y_i - \bar{y})^2}{n-1}$

Two-phase stratification, $\hat{Y}_2 = N\sum_{h=1}^{3} w_h \bar{y}_h = N\sum_{h=1}^{3} \dfrac{n'_h}{n'} \dfrac{\sum_{i=1}^{n_h} Y_{hi}}{n_h}$;

$$v(\hat{Y}_2) = N^2\left[\sum_{h=1}^{3} \frac{w_h^2 s_h^2(y)}{n_h} - \sum_{h=1}^{3} \frac{w_h s_h^2(y)}{N} + \frac{(N-n')}{(N-1)n'} \sum_{h=1}^{3} w_h \left(\bar{y}_h - \frac{\hat{Y}_2}{N}\right)^2\right],$$

where $s_h^2(y) = \dfrac{\sum_{i=1}^{n}(Y_{hi} - \bar{y}_h)^2}{n_h - 1}$

Two-phase ratio, $\hat{Y}_3 = \sum_{c=1}^{5} \hat{M}^c \bar{y}^c = \sum_{c=1}^{C} \left(N\dfrac{\sum_{i=1}^{n'} M_i^c}{n'}\right)\left(\dfrac{\sum_{i=1}^{n} Y_i^c}{\sum_{i=1}^{n} M_i^c}\right)$;

$$v(\hat{Y}_3) = N^2\left[\left(\frac{1}{n'} - \frac{1}{n}\right) s^2(y) + \left(\frac{1}{n} - \frac{1}{n'}\right) s^2(d)\right],$$

where $s^2(d) = \dfrac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{n-1}$; $Y_i^* = \sum_{c=1}^{C} M_i^c \bar{y}^c$

FIGURE 1

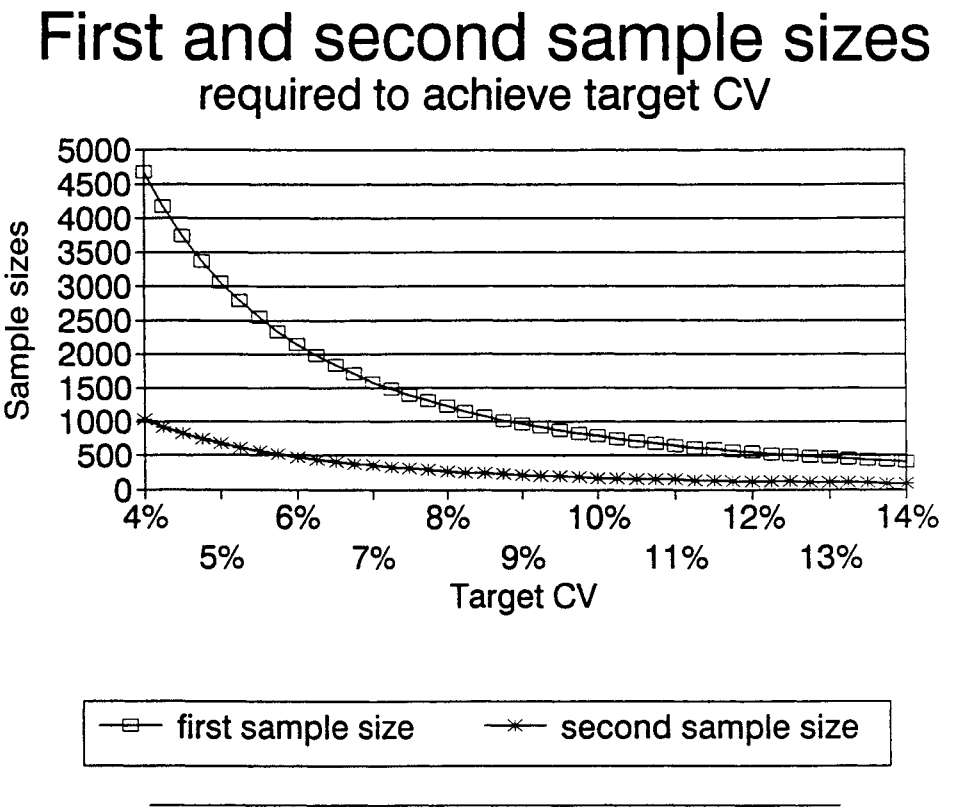

TABLE 1. CLASSIFICATION AND VERIFICATION OF PAIRS OF HOUSEHOLDS

CLASS (C)		NUMBER OF PAIRS PER EA ($\bar{m}^c$)	INCIDENCE OF OVERCOVERAGE (I^c)
1	MORE THAN ONE SIMILAR PERSON SAME NEIGHBOURHOOD	0.10	0.94
2	MORE THAN ONE SIMILAR PERSON OUTSIDE NEIGHBOURHOOD	0.36	0.35
3	SIMILAR SINGLE-PERSON HHLD SAME NEIGHBOURHOOD	0.13	0.50
4	SIMILAR SINGLE-PERSON HHLD OUTSIDE NEIGHBOURHOOD	1.85	0.02
5	ONLY ONE SIMILAR PERSON SAME NEIGHBOURHOOD	1.50	0.01
6	ONLY ONE SIMILAR PERSON OUTSIDE NEIGHBOURHOOD	146	(-)
7	NO SIMILAR PERSON	30646	(-)

TABLE 2. CHARACTERISTICS OF THE POPULATION USED FOR SIMULATING THE THREE SAMPLE DESIGNS

2.1 Frequency distribution of within EA overcoverage

Y_i	0	1	2	3	4	5	6	7	8	11
Frequency	284	20	22	18	16	4	7	5	3	1

2.2 Population statistics

Statistic	Stratum level h=1	h=2	h=3	All
N	31	77	272	380
Y	141	142	31	314
$S^2(y)$	5.922	3.949	0.190	3.104
v	0.6	0.5	0.1	

Statistic	Class level c=1	c=2	c=3	c=4	c=5
M	40	138	50	705	572
Y	127	137	22	17	11
I	0.93	0.35	0.44	0.02	0.01
$S^2(d)$ = 1.051					

TABLE 3. RESULTS OF THE SIMULATION OF THE THREE SAMPLE DESIGNS

Initial sample of n' = 80

STATISTIC	SRS	TP_STR	TP_RAT
average sample (n)	24.5	19	
average estimate	306	318	302
average estim. coeff. of var	43.9 %	28.9 %	34.1 %
observed coverage rate	.90	.90	.78
confidence interval too low	.09	.09	.17
confidence interval too high	.01	.01	.05

n' = 120

SRS	TP_STR	TP_RAT
36	28	
321	315	306
34.3 %	22.8 %	26.5 %
.94	.92	.82
.06	.07	.13
.00	.01	.05

Initial sample of n' = 160

STATISTIC	SRS	TP_STR	TP_RAT
average sample (n)	48	37	
average estimate	313	314	320
average estim. coeff. var.	29.4 %	18.8 %	21.7 %
observed coverage rate	.91	.93	.87
confidence interval too low	.08	.06	.09
confidence interval too high	.01	.01	.04

n' = 200

SRS	TP_STR	TP_RAT
59	45	
308	311	318
25.8 %	16.2 %	18.5 %
.91	.92	.89
.08	.07	.07
.01	.01	.04