THE 1996 AUTOMATED MATCH STUDY

Michael Mayda, Claude Julien, Statistics Canada
Michael Mayda, Statistics Canada, RH Coats Bldg 15-C, Ottawa ON K1A 0T6 e-mail: maydmic@statcan.ca

Key Words: record linkage, automated matching, census coverage error measurement, overcoverage

1. Introduction

For the 1996 Canadian Census of Population, several studies will be conducted to measure coverage error, which occurs when persons, households or dwellings are either missed by the census or enumerated in error. In this paper, we focus on the methodology of Statistics Canada's main overcoverage measurement study, the Automated Match Study (AMS), currently under development.

Section 2 gives a general background to the problem of census overcoverage, and how it is measured by Statistics Canada, both by the AMS and the other major coverage study, the Reverse Record Check (RRC). Section 3 highlights some of the advantages and disadvantages of the AMS as compared to the RRC. In Section 4, we present results obtained from the 1991 AMS and a pilot study conducted using 1991 census data, and show how these results are being used to design the 1996 study. Section 5 outlines the work still underway and points to directions that the work may take in the future.

2. General Background

In the 1991 Census, overcoverage was estimated at 0.6%. Although this may appear to be a small figure, accurate measurement of overcoverage is important nonetheless because, in Canada, federal government transfers to the provinces are based on census figures that are adjusted for net undercoverage error (undercoverage less overcoverage).

There are two types of overcoverage: duplicate enumeration of the same person, and enumeration of persons not in the target population (ficticious persons, deceased persons, pets, foreign visitors, etc.). The latter type is rare, estimated at less than 0.1% in 1991. Consequently, it is not vigorously pursued by the overcoverage studies. The former can be caused by factors related to the respondent (moving close to census day, maintaining more than one residence, etc.), or it can be caused by procedural errors (delivering two questionnaires to the same dwelling, capture of the same questionnaire twice, etc.).

At Statistics Canada, estimates of overcoverage are produced by combining the results of several studies. This is done to take advantage of the relative efficiencies of various methodologies, but also because no study alone is capable of measuring all types and causes of overcoverage. The two main overcoverage studies are described below. See Statistics Canada (1994) for details about the complete 1991 Coverage Error Measurement Program.

The AMS

At the heart of the AMS lie computer matching programs which identify pairs of households that are "similar". Similarity is described in terms of the number of persons matched between households, the sizes of the two households and their relative proximity to each other. Since names are not present on the census database, persons are matched on sex and date of birth. Two persons with the same sex and day, month and year of birth are said to exactly match. If three of the four components are the same, or just the day and month are inverted, persons are said to nearly match. For further details about the matching programs, see Julien and Mayda (1995).

Once pairs of similar households are detected, the census questionnaires for a sample of them will be manually verified to determine how much overcoverage, if any, occurred. This determination will be based on the names found on the questionnaires. Naturally, many households will appear similar to another due to chance. However, intuitively, and as we will see in Section 4, the likelihood that a pair of similar households contains overcoverage increases dramatically as the similarity increases.

The RRC

Although primarily an undercoverage study, this study measures some overcoverage. Prior to census day, a sample of persons who should be enumerated is selected. This sample is drawn from a number of list frames including: the previous census, intercensal births and immigration, and persons missed by the previous census. As no complete frame for the last group exists, a sample of such persons as determined by the previous RRC is used. Shortly after the census is complete, the selected persons are traced and interviewed in order to obtain all possible addresses at which they may have been enumerated. A searching operation will find the census

questionnaires completed at these addresses and determine how many times the selected persons were enumerated (once, twice, not at all).

3. Advantages and disadvantages of the AMS

The AMS can measure most duplicate enumerations, whether caused by procedural errors or factors related to the respondents. The RRC primarily measures duplicate enumerations due to the respondents. For example, a respondent to the RRC would be unaware that his census questionnaire was captured twice.

The AMS does not measure enumeration of persons not in the target population, and the RRC only some of it. As stated at the outset, this is a minor overcoverage source and does not present a great difficulty.

One disadvantage of the AMS concerns small, and especially single person households. This is because the study's success rests on the fact that as a household's size increases, the chance that it is unique when viewed as the collection of the sex and date of birth information of its members, increases.

Another potential drawback is the reliance on the quality of the sex and birth date data. The study may suffer if the quality is poor, but fortunately this has not been the Canadian experience.

Despite its drawbacks, as we will show, the AMS will provide estimates of relatively high precision when compared to the RRC. Moreover, the AMS can be conducted relatively cheaply because it does not require the expensive telephone tracing and interview process which characterizes the RRC.

4. 1991 results

From Canada's population of ten million households, one set of matching programs identified 280,000 pairs of households on the 1991 census database with at least two exact matches (first three rows of Table 1).

A second set of programs was applied to a sample of EAs, in order to estimate the number of pairs having at least one near match, but less than two exact matches (rows four through eight of Table 1). For 1996, though, we will apply these programs to every EA in order to determine the true population distribution.

The three proximities represent successively wider areas. Namely:

1. pairs within Enumeration Areas (EAs);
2. pairs within Federal Electoral Districts (FEDs), but in different EAs; and,
3. pairs within provinces, but different FEDs.

To put these in context, EAs average 600 persons and FEDs, 90,000. Provinces range in size from 130,000 to 10 million.

The 1991 AMS was limited to identifying and verifying pairs of similar households within EAs only.

Table 1 Population distribution of pairs of similar households (N)

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	2208	1575	1046
2	E=3	2701	2103	1553
3	E=2	6022	10369	251220
4	E=1, N≥1	8248	-	-
5	E=1, N=0	275525	-	-
6	E=0, N≥2	236506	-	-
7	E=0, N=1, $S_1=S_2=1$	468554	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	26670187	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household

Table 2 Sample distribution of pairs of similar households (n)

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	21	67	15
2	E=3	34	115	66
3	E=2	222	337	240
4	E=1, N≥1	177	-	-
5	E=1, N=0	2112	-	-
6	E=0, N≥2	2952	-	-
7	E=0, N=1, $S_1=S_2=1$	4310	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	4509	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household

However, recent improvements to our computer matching algorithms enabled us to carry out a pilot study in which the more widely separated pairs were identified and a sample of them manually verified for overcoverage. Together, these two sources provided over 15,000 verified pairs for us to analyze. Table 2 gives their distribution. For further details of the 1991 AMS and the pilot study, see Reedman (1993) and Bernier (1995), respectively.

The variable of interest is the number of persons overcovered in a given pair. Table 3 displays the averages observed per pair. The estimated variances of the overcoverage per pair are found in Table 4.

The percentage of pairs in each cell having any overcoverage is shown in Table 5. The shaded portions of Table 5 indicate where the percentages are well above 90%, and, in fact, are 100% in a few cases. The technique has, therefore, identified pockets where the incidence of overcoverage is very high. Considering the scarce nature of overcoverage globally, this will prove to be a useful achievement. On the down side, the utility of strata with no exact matches is limited. This is shown in rows six through eight of Table 5.

By multiplying the entries of Table 1 with the corresponding entries of Table 3, estimates of the level of overcoverage in each cell were obtained and are presented in Table 6. Notably, the shaded cells in Table 6 account for 56,000 overcovered persons. To put this into perspective, this represents 35% of the 160,000 persons estimated to have been overcovered in all of Canada in 1991. That there are only 17,000 pairs of households involved (as shown in Table 1) demonstrates the methodology's effectiveness.

By the new AMS methodology, overcoverage in 1991 was estimated at approximately 100,000 persons (total of the entries in Table 6). This represents about 65% of the total as estimated by the 1991 overcoverage studies combined.

For 1996, a sample of 4500 pairs is planned. To assess the performance possible, the Neyman allocation was determined and is shown in Table 7. This allocation is for illustration only, since the actual sample will be allocated by province in order to achieve adequate precision at the provincial level. It is important to note that while the AMS methodology cannot detect all overcoverage, what it can measure, it does with good precision. With the allocation of Table 7, the estimated co-efficient of variation for the estimated total is just 3.3%. In contrast, a simple random sample (which approximates the efficiency of the RRC) of about 250,000 persons would be needed to achieve the same level of precision.

Table 3 Average number of overcovered persons per pair ($\bar{y}$)

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	4.674	4.417	4.120
2	E=3	3.127	3.317	3.106
3	E=2	2.292	1.411	0.026
4	E= 1, N≥1	1.150	-	-
5	E=1, N=0	0.044	-	-
6	E=0, N≥2	0.008	-	-
7	E=0, N=1, $S_1=S_2=1$	0.002	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	0.000	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household

Table 4 Variance (s^2)

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	0.761	0.099	0.009
2	E=3	1.000	0.341	0.365
3	E=2	0.568	1.437	0.065
4	E=1, N≥1	2.135	-	-
5	E=1, N=0	0.068	-	-
6	E=0, N≥2	0.021	-	-
7	E=0, N=1, $S_1=S_2=1$	0.002	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	0.000	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household

5. Current research and future directions

Although effective, the stratification outlined here can be improved. In particular, we are looking into taking more account of household size. For

example, two five person households with only two matches between them are less likely to represent overcoverage than two two person households with two matches between them. Furthermore, it is possible for overcoverage to occur in excess of that indicated by the number of matched persons in a given pair of households. This can happen when response errors in the census sex and birth date data prevent the matching algorithms from detecting a match. The implication is that even in strata where 100% of the pairs contain some overcoverage, the variation in the overcoverage per pair can be reduced if strata are constructed with better control of the household sizes.

We are also investigating household pairs located in different provinces. The limited data we have (not presented here) suggest that the "stronger" strata are still viable, but that little overcoverage actually occurs between provinces.

We are also assessing the suitability of creating strata with less than two exact matches between the paired households and where the two households lie in different EAs. At present, direct survey estimates for such strata are unavailable, so we are looking into how we might model estimates for these strata. To outline how we might do this, consider, for example, that from Table 6, among households with 2 exact matches (row 3), approximately equal numbers of overcoverage cases occur within EA as within FEDs. It might be reasonable to assume that this pattern also holds in row 4. If so, then there are approximately 10,000 overcovered persons in this cell. Assuming each pair contains either no overcoverage or two overcovered persons implies that approximately 5000 pairs (10,000/2) contain overcoverage. The matching programs can be readily adapted to determine the total number of pairs in this cell, and the average per pair and the requires variance derived. This procedure should provide reasonable advance estimates upon which to base the 1996 allocation.

Also of interest is determining the point at which we may want to exclude strata from the AMS. Since some of the strata shown here have such small overcoverage rates, we may be better off using the RRC.

The occurrence of a given household in more than one household pair can pose problems. For example, suppose a person was counted by the census in households A, B and C. If the three households were similar to one another, we would generate the pairs (A,B), (A,C) and (B,C). If we were to independently verify each of these pairs, we would incorrectly conclude that there were three instances of overcoverage, when in fact there are only two. The little data we have suggests, however, that triple coverage is rare, so this is not likely to be a great problem. However, if no overcoverage were actually

Table 5 Percentage of pairs with overcoverage

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	100.0	100.0	100.0
2	E=3	93.2	100.0	97.8
3	E=2	98.1	62.3	1.1
4	E=1, N≥1	40.1	-	-
5	E=1, N=0	3.5	-	-
6	E=0, N≥2	0.3	-	-
7	E=0, N=1, $S_1=S_2=1$	0.2	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	0.0	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household

Table 6 Estimated overcoverage ($\hat{y}$)

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	10321	6957	4309
2	E=3	8446	6975	4824
3	E=2	13803	14632	6427
4	E=1, N≥1	9485	-	-
5	E=1, N=0	12137	-	-
6	E=0, N≥2	1875	-	-
7	E=0, N=1, $S_1=S_2=1$	761	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	0	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household

present and the households were similar due to chance, it becomes an efficiency concern since the matching programs would have created more household pairs than necessary.

In the long term, since many "obvious" cases of overcoverage can be detected, in future censuses consideration could be given towards removing them from the census database entirely.

The results indicate that record linkage of persons based on the sexes and dates of birth of their household members may be useful in a wide range of applications. For example, Julien and Mayda (1995) outline how the match methodology will be used to link 1996 RRC survey data to the 1996 census database.

6. Conclusion

The 1996 Automated Match Study will make extensive use of computer matching to estimate overcoverage in the 1996 Census. It will be relatively inexpensive to conduct, and although it cannot measure all types of overcoverage, what it can measure, it does with far greater precision than alternative methodologies.

7. Acknowledgements

We wish to acknowledge Julie Bernier whose work on the pilot study forms the basis for many of these results. We would also like to thank our reviewers who provided us with useful comments.

8. References

Bernier, J. (1995). Étude pilote: estimation du surdénombrement détecté par appariement automatique. Unpublished paper, Social Survey Methods Division, Statistics Canada.

Julien C., and Mayda M. (1995). Improving census coverage error measurement through automated matching. *American Statistical Society 1995 Proceedings of the Section on Survey Research Methods*, Volume II, 849-854.

Reedman, L. (1993). Automated Match Study methodology report. Unpublished paper, Social Survey Methods Division, Statistics Canada.

Statistics Canada (1994). Coverage - 1991 Census Technical Reports. Catalogue 92-314E, Statistics Canada.

Table 7 Trial Neyman allocation

	MATCH TYPE	PROXIMITY		
		EA	FED	PROV
1	E≥4	38	10	2
2	E=3	54	25	19
3	E=2	91	248	1274
4	E=1, N≥1	241	-	-
5	E=1, N=0	1439	-	-
6	E=0, N≥2	683	-	-
7	E=0, N=1, $S_1=S_2=1$	377	-	-
8	E=0, N=1, S_1 or $S_2\geq 2$	0	-	-

E = Exact matches, N = Near matches,
S_1, S_2 = Size of each household