IMPROVING CENSUS COVERAGE ERROR MEASUREMENT THROUGH AUTOMATED MATCHING

Claude Julien and Michael Mayda, Statistics Canada
Claude Julien, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6

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

1. Introduction

Record linkage may be described as the process of comparing files of records in order to identify pairs of records that relate to the same population unit. In this paper, we will discuss several applications of record linkage to the improvement of census coverage and to the measurement of census coverage errors. We will examine linking census records together in order to identify, and possibly remove, population units which are present more than once, and hence represent census overcoverage. We will also examine linking data from a survey of persons to a census database in order to determine their census enumeration status (ie. missed or enumerated).

Ideally, record linkage of persons is carried out on unique identifiers (e.g. Social Insurance Number), but in practice it often must rely on non-unique identifiers such as names, addresses and demographic characteristics (ie. sex, date of birth). In the context of the Canadian Census, record linkage is even more difficult. To help ensure the confidentiality of respondents' data, names and addresses are absent from the database, and are therefore unavailable for use as matching variables. Moreover, the sheer size of the database imposes a heavy computational burden and limits the types of matching that can be attempted.

In designing the 1996 Canadian Census Coverage Studies, we developed and evaluated a matching algorithm that identifies all pairs of "similar" households on a given database, using only the sex and date of birth of the household members. In this context, similar households (also referred to as links or matches) have at least two members in common; that is, two members with the same sex and date of birth. The algorithm is described in Section 2.

The algorithm was evaluated in two ways. In Section 3, we show how it can be used to improve the measurement of census undercoverage. In Section 4, we assess the potential of using this algorithm to measure census overcoverage, and to a further extent, to eliminate some of it. Section 5 concludes with a description of future work.

2. The matching algorithm

The purpose of the matching algorithm is to identify pairs of similar households on a database. As input, the process requires a person level file containing the household identifier, SEX and Date Of Birth (SEXDOB) for each person. The full date of birth is required (year, month and day). Using the given household identifiers, the process creates household rosters and selects only those with two or more persons with complete and valid SEXDOBs. Hereafter, we shall refer to these as matchable households. Since matchable households contain at least two members, it should be noted that single person households cannot be matched by this procedure. The matching algorithm is straightforward and is summarized by the following four steps:

Step 1 For each matchable household, from the persons within, generate all possible combinations of two persons (couples). Each household will generate $\binom{n_i}{2}$ couples, where n_i is the number of persons with complete and valid SEXDOBs in the ith household. For a given couple, concatenate the two SEXDOBs such that the youngest female is first and the oldest male last (the actual order is arbitrary, but must be consistent), and create a file, one record per couple, containing the household identifier and the concatenated string. For example, the following household:
Male 1962/04/10, Female 1963/08/04 and Female 1990/04/01
generates the following strings:
F19900401F19630804, F19900401M19620410, F19630804M19620410.

Step 2 From the above file, delete any record whose string is unique to a single household.

Step 3 Group the remaining records by string. For each string, generate all possible combinations of pairs of household identifiers sharing that string. For each pair of households, arbitrarily define HHLD1 as the one with the smaller household identifier and HHLD2 as the one with the larger identifier. Create a file, one record per pair, containing HHLD1 and HHLD2.

Step 4 From the above file, keep only one record per combination of HHLD1 and HHLD2 (a given pair of households is duplicated for each couple that the households have in common). Furthermore, any pair where HHLD1 = HHLD2 should be removed (this can happen if couples are not unique within a particular household, as, for example, could occur if a household contains a set of twins).

Step 4 creates a file of all pairs of households having at least two persons in common. Each pair is then compared in more detail and categorized based on the number of persons who exactly match (sex, day month and year of birth are the same), the number of persons who nearly match (three of sex, day, month and year are the same), and the number of persons in each of the two households.

The algorithm was programmed in PL-1 and runs on Statistics Canada's mainframe. Four databases were created by appending the survey data from the Reverse Record Check Study (RRC) (described in the following section), to each of the four 1991 Census regional data bases. A total of 10 million households were processed by the algorithm, of which 7.5 million were matchable. Although Step 1 of the algorithm generated 31.5 million couples, only 0.5 million non-unique couples were identified during Step 2. Finally, in Step 4, the algorithm found 81,330 survey households similar to census households (survey-census matches), and 355,905 pairs of similar census households (census-census matches). Total processing time, for all four databases, was less than 20 minutes CPU.

In the next two sections, we will analyze the survey to census and census to census links.

3. Linking survey data to the Census

The RRC is the main study that measures undercoverage in the Canadian Census [1]. A sample of persons who should be enumerated is selected from sources independent of the current census. The sources include: the previous census, registration of intercensal births, landed immigrants, persons in Canada on special permits, refugee claimants, and a sample of persons who were missed in the previous census. Tracing and data collection operations are carried out to obtain all addresses where a selected person (SP) may be enumerated. Information about household members living with the SP is also collected. A combination of automated and clerical operations are carried out to geocode addresses and to find the census questionnaire completed at each address. The information on the questionnaire determines if a SP is enumerated or missed.

In 1991, a sample of 55,912 persons was selected, of which 48,227 persons were classified as enumerated and coded to a census household identifier (CENID) composed of Province (PROV), Federal Electoral District (FED), Enumeration Area (EA) and Household Number (HHLD). The remainder of the sample was either missed, out of scope (ie. deceased prior to census day), or the enumeration status could not be determined.

Of the survey to census matches, we kept only those for which the SP himself was either exactly or nearly matched. The resulting 44,195 matches were classified into one of the following categories:

Class	**Description**
A	At least three persons match exactly
B	Two persons match exactly and at least two persons nearly match
C	Two persons match exactly and both households have only two persons
D	Two persons match exactly and only one person nearly matches
E	Two persons match exactly and both households have four persons or less
F	Two persons match exactly and both households have five persons or more

We further classified the matches into one of the following categories representing the geographic proximity of the census household CENID to the RRC household CENID:

Proximity	**Description**
0	Same PROV, FED, EA and HHLD (ie. same CENID)
1	Same PROV, FED and EA
2	Same PROV and FED
3	Same PROV
4	Same region (Atlantic, Quebec, Ontario or West)
5	Different region

On average, EAs contain approximately 600 persons, and FEDs about 90,000. The provinces range in size from 125,000 to 10,000,000, and the four regions range in size from 2,000,000 to 10,000,000. There were very few strong matches occurring between different provinces. In this paper, we limit the analysis to those matches occurring within the same province (Proximity < = 3).

The distribution of matches is presented in Table 1. The 34,903 matches occurring at the CENID are true matches. That is, the algorithm matched the RRC household to the census household where the SP was found enumerated by the regular RRC processing. This means that the matching algorithm linked (or found) 72% of the 48,227 enumerated SPs to the households they were coded to. The algorithm did not match the other enumerated SPs for a variety of reasons, including: SPs enumerated in single person households, or response errors in either the RRC households or the census households.

Most of the A and B-type matches produced by the algorithm occur at the CENID. This suggests that A and B-type matches are very likely to be true matches. On the other hand, a fair proportion of E and F-type matches occur in another FED within the same province. This would indicate that they are unlikely to be true matches.

Table 1 - Distribution of survey-census matches

	Proximity				
Class	0	1	2	3	Total
A	23,283	46	96	62	23,487
B	725	1	9	6	741
C	5,552	13	15	62	5,642
D	1,794	7	13	53*	1,867
E	2,993	4	32	462*	3,491
F	556	0	26	788*	1,370
Total	34,903	7	191	1,433	36,598

* not verified

Matches that did not occur at the CENID were potentially false (ie. the census household linked to the RRC household did not contain the SP). All potential matches occurring within the same FED were verified by comparing the RRC household to the persons listed on the census questionnaire to determine whether the SP was enumerated (true match) or not (false match). In addition, A, B and C-type matches occurring in Proximity=3 were verified. D, E and F-type matches occurring in Proximity=3 were not verified, and were assumed false. The distribution of false matches is presented in Table 2 and the likelihood of a true match by Class and Proximity is presented in Table 3. The likelihood of a true match is the ratio of the number of matches minus the false matches to the number of matches.

Table 2 - Distribution of false survey-census matches

	Proximity			
Class	1	2	3	Total
A	0	0	1	1
B	0	0	2	2
C	0	0	42	42
D	0	3	53*	56
E	0	10	462*	472
F	0	15	788*	803

* assumed false

Table 3 - Likelihood of a true survey-census match

	Proximity			
Class	1	2	3	Total
A	1.00	1.00	0.98	1.00
B	1.00	1.00	0.67	1.00
C	1.00	1.00	0.32	0.99
D	1.00	0.77	0.00*	0.97
E	1.00	0.69	0.00*	0.86
F	1.00*	0.42	0.00*	0.41

* assumed rate

Entries in the Total column of Tables 1, 2 and 3 show, respectively, that there are 23,487 A-type matches and only 1 false match, yielding a likelihood rate practically equal to 1.00. This means that a census household almost certainly contains the SP when it satisfies the following conditions: it is located in the same province as the SP, it has at least three persons in common with the SP's RRC household, and the SP himself is either nearly or exactly matched. In fact, based on our observations, all persons in a RRC household who are involved in an A-type match (exact or near match) are the same persons as on the census file.

Another interesting result is that all matches occurring within an EA are true. This implies that all couples (any combination of two persons living in the same household) within an EA are unique. Furthermore, within FEDs, all A, B and C-type matches are true.

In the 1996 RRC, when a person's enumeration status will have to be determined, the algorithm will be used to complement the time-consuming manual procedures used to search for the census questionnaires completed at the addresses obtained during tracing. It will be used to quickly find the "easy" cases and allow more resources to be used to search for the more difficult cases, such as single person households. Every address will be geocoded to a search area (set of EAs) no bigger than a FED. It is clear that all A, B and C-type matches observed in a search area will confirm that the SP is enumerated, without having to look at the census questionnaire, and with negligible chance of misclassifying the SP as enumerated.

The matching algorithm will also provide a list of census households, located anywhere in Canada, where a SP may be enumerated. In this case, all matches that do not occur in a search area will be verified to determine whether the SP is enumerated or not. This will allow us to find persons easily even when a geocoding error identifies the wrong search area for the address.

Based on our analysis, the matching algorithm will be useful, in one way or another, in finding approximately 75% of all enumerated SPs.

4. Within Census record linkage

The 1991 Canadian Census of Population marked the first time a complete study was conducted to estimate census overcoverage. There are two types of overcoverage: persons enumerated more than once and erroneous enumeration (foreign visitors, fictitious persons, etc.). The 1991 Overcoverage Study [1] consisted of three components, and one of these components, called the Automated Match Study (AMS), estimated the number of persons enumerated more than once within the same EA.

A program was developed in 1991 which compared each household with each other household in a given EA, in order to identify those that had persons in common. In an EA of N households, N * (N-1) / 2 comparisons were done. Given the computational burden of this procedure, the scope of the AMS was limited to EAs. Nonetheless, the 1991 AMS produced an estimate of over 44,000 persons double-counted within an EA. This estimate represents 25% of all overcoverage in the 1991 Census.

In this section, we analyze the matches between households on the census database as a means of detecting overcoverage caused by duplicate enumeration, and show how our algorithm will be used to design an improved AMS in 1996.

As mentioned in Section 2, 355,905 pairs of similar households were identified by the matching algorithm. Like the analysis of the RRC data, we will limit the analysis to the matches occurring within the same province. The pairs were classified by strength of the match and the geographic proximity of the two census households. This distribution is presented in Table 4. Unlike the analysis of survey to census matches, there are no entries for Proximity=0 since census households trivially match to themselves.

Table 4 - Distribution of census-census matches

	Proximity			
Class	1	2	3	Total
A	4,909	3,679	2,640	11,228
B	360	259	262	881
C	3,486	2,422	8,888	14,796
D	951	859	6,290	8,100
E	842	3,183	87,898	91,923
F	387	3,662	148,199	152,248
Total	10,935	14,064	254,177	279,176

The distribution of matches by Class and Proximity is consistent with the results observed in the RRC application and the limited knowledge we have on overcoverage. For example, strong matches like A-types, are more frequent within the same EA (Proximity=1) than within the same FED (Proximity=2), and more frequent within FEDs than within provinces (Proximity=3). In contrast, the number of E and F-type matches increases rapidly as the distance between the households increases. This indicates that the likelihood of a true match decreases as the distance increases.

Table 5 presents the number of persons matched (nearly or exactly) within the similar pairs of households. By multiplying these numbers with the estimated true match rates in Table 3, we can project a total of 79,299 persons who are enumerated more than once. These projections, presented in Table 6, are crude and understate the total level of overcoverage. In fact, the 1991 Overcoverage Study estimated that more than 159,000 persons were overcovered in the 1991 Census. Several factors account for the difference between the two figures. First of all, some duplication cannot be detected by this matching algorithm (ie. census households with only one person in common). Secondly, because of response errors, some duplication cannot be detected or some pairs of similar households have more persons in common than the number of persons who match would indicate. Thirdly, overcoverage due to erroneous enumeration (5% of all overcoverage) cannot be detected.

Table 5 - Persons involved in the census-census matches

	Proximity			
Class	1	2	3	Total
A	19,424	14,344	10,046	43,814
B	1,513	1,090	1,088	3,691
C	6,972	4,844	17,776	29,592
D	2,853	2,577	18,870	24,300
E	1,684	6,366	175,796	183,846
F	774	7,324	296,398	304,496
Total	33,220	36,545	519,974	589,739

The algorithm will form the basis of the design of the 1996 AMS, which will be the key component for measuring overcoverage in the 1996 Census Coverage Error Measurement Program (other studies, will measure the overcoverage that cannot be detected by the AMS). The pairs of similar households will be stratified by Class and Proximity as presented in Table 4. A sample of pairs will be selected from each stratum and their names on census questionnaires will be verified to determine the number of persons enumerated more than once. A pilot study is currently underway in which a sample of pairs of households from the 1991 Census database is being processed. This study will produce more reliable estimates of the true match rates and expected overcoverage than those presented here. These parameters will then be used as input into the design of the 1996 AMS.

In the longer term, the results of the 1996 AMS may confirm that some strata contain only true matches and we could use the matching algorithm to identify obvious cases of overcoverage and automatically remove them from the census database.

Table 6 - Projection of overcoverage (multiple counts)

	Proximity			
Class	1	2	3	Total
A	19,424	14,344	9,884	43,652
B	1,513	1,090	725	3,328
C	6,972	4,844	5,734	17,550
D	2,853	1,982	0	4,835
E	1,684	4,377	0	6,061
F	774	3,099	0	3,873
Total	33,220	29,736	16,343	79,299

Based on the limited results known so far, 63,382 persons could have been removed from the 1991 census database. This is the number of persons involved in: all matches occurring within an EA; A, B and C-type matches occurring with a FED; and, A-type matches occurring within a province.

5. Conclusion and future work

In designing the 1996 Canadian Census Coverage Measurement Studies, we have developed an algorithm which links households based only on readily available demographic characteristics of the household members.

In the 1996 Census, the algorithm will be used to improve the measurement of coverage error. It will be an integral part of the process that determines the census enumeration status of persons selected from sources independent of the census. Our studies indicate that the algorithm can correctly link between 65% and 75% of all households provided by the external sources to the census database. This ability will be employed to classify many selected persons as enumerated, using relatively little resources. The remaining resources would then be concentrated on assigning the enumeration status to persons who cannot easily be classified.

The algorithm will also be used to design a study that will measure more than half of all overcoverage. Moreover, this half will be measured with very high precision. We are currently conducting a pilot study, based on 1991 Census data, that will provide the required parameters to design the 1996 AMS.

In terms of coverage improvement, the algorithm has the potential to remove more than a third of all census overcoverage. This potential will be assessed extensively and, if confirmed, the algorithm could be implemented to remove overcoverage directly from the database in the 2001 Census.

Finally, only one external data source was studied: survey data from the RRC. The algorithm could be used to link any external source, as long as it has the required household identifiers and sex and date of birth of the household members.

References

[1] Statistics Canada, Coverage - 1991 Census Technical Reports, Catalogue 92-314E, 1994