ADJUSTING FOR NON-RESPONSE IN THE 1996 REVERSE RECORD CHECK

Martin Provost, Statistics Canada
R.H. Coats Building, 16th Floor, Tunney's Pasture, Ottawa ONT, Canada, K1A 0T6

Key Words: coverage error, Census, non-response adjustment, weight adjustment group, undercoverage

1. Introduction

The Reverse Record Check (RRC) is one of three studies designed to assess the coverage of the Canadian Census. The last Census of population took place on May 14th 1996. The accuracy of the results of the coverage studies are of critical importance because, since 1991, transfer payments from the federal to the provincial governments are calculated using the Census count adjusted for coverage errors, instead of the plain Census count. Because the non-response adjustment has a non-negligible impact on the final estimates of coverage error, the methodology used for this adjustment has high user visibility.

We will see in this paper how the non-response adjustment was done for the preliminary results released in March 1998, and the hypotheses behind that adjustment. We will also describe and justify the changes that have since been introduced in the methodology along with the impact of the chosen adjustment on the results of the coverage studies.

2. Census Coverage Studies

There are two types of coverage error that occur with the Census. The first, undercoverage, occurs when an individual that belongs to the Census universe is missed by the Census, i.e. not included on any Census questionnaire. The second coverage error, overcoverage, occurs when an individual is counted more than once, or when an individual that doesn't belong to the Census universe (e.g. deceased person, foreign visitor, etc.) is included on a questionnaire.

Three studies are designed to measure the coverage of the Census of population. The first is the Reverse Record Check Study, which measures all the undercoverage and a part of the overcoverage. The two other studies, the Collective Dwelling Study and the Automated Match Study, measure specific types of overcoverage. In this paper, we will focus on the RRC, and particularly on the measurement of undercoverage. According to the preliminary results, the undercoverage rate was 3.33% while the overcoverage rate was 0.76%, leaving a net undercoverage rate of 2.57%.

3. The Reverse Record Check Study

The 1996 RRC consists of a sample of 57,016 persons selected from sources independent of the current Census. For each person in the sample, an attempt is made to determine whether or not they were enumerated by the Census, and, if so, how many times. The sample is selected from six different sources that, once combined, cover the Census target universe. These frames are the following:

1) Census Frame: Persons enumerated in the 1991 Census.
2) Birth Frame: Babies born in the intercensal period.
3) Immigrant Frame: Immigrants that arrived in Canada in the intercensal period.
4) Missed Frame: Persons missed by the 1991 Census.
5) Non Permanent Resident (NPR) Frame: Persons that were in Canada on Census day with either a work permit, a student permit or a ministerial permit plus the refugee claimants (since 1991, these persons are part of the Census population).
6) Health Care Files (HCF): Persons covered by the health care plan of one of the two territories.

The first five frames cover the population of the ten provinces, while the sixth frame covers the two territories, the Yukon and the Northwest Territories (The sample for the two territories was selected exclusively from this frame.). Because we don't have a list of all the persons missed by the previous Census, the missed frame is only a conceptual frame. It contains all of the 2,341 persons that were identified as missed by the 1991 RRC. All of these persons were selected for the 1996 sample. The five other frames were stratified by province and other characteristics like demographic information for the Census frame and year of birth for the birth frame. A sample was then selected from each stratum.

For each Selected Person (SP) included in the sample and every member of their selection household (people that were living with the SP according to the frame information), we tried to update the address information by linking to administrative files. Staff in our regional offices then attempted to trace every SP. When they could trace an SP, they proceeded with a telephone interview. During the interview, the SPs were asked for their Census day address, the name, sex and date of birth of every person that was living with them at that time, and any other address where they could have been included on a Census questionnaire.

Following tracing and collection, every address was processed and the associated Census Form was verified.

Each SP was assigned a classification indicating:
1) Enumerated once / more than once;
2) Missed by the Census;
3) Deceased before Census day;
4) Emigrated before Census day;
5) Abroad at the time of Census; or
6) Out of scope / frame overlap.

For some of the SPs, it was not possible to assign a classification. These are the non-respondents. The next section describes the non-respondents.

4. Description of the Non-Response

There are three types of non-response in the Reverse Record Check. The first type is the Not Identifiable (NI) persons. These are SPs for whom the selection information is insufficient to even attempt tracing. Most of the Not Identifiable comes from Form 4s. Form 4s are filled when the Census enumerator encounters an absent household which he/she thinks is not vacant. Household size and demographic characteristics are later imputed (In some cases, the enumerator knows the household size, so only the characteristics of the household members are imputed.). Sometimes when we select someone from the 1991 Census frame, we go to the Census questionnaire to collect the selection information for the SP and we realise that the selected address corresponds to a Form 4. We then classify the SP as Not Identifiable. Form 4s account for 400 of the 440 Not Identifiable. The others are cases for which the selection information is too vague to be useful for tracing.

The remaining SPs are sent for tracing. In some cases, all attempts to trace the person fail. The SP is then classified as Not Traced (NT). That's the second type of non-response. The third level of non-response consists of the Not Classified (NC) persons. These are SPs that have been traced successfully but for whom we can't assign a final classification. For these cases, we know that the person was enumerable, i.e. they should have been included on a Census questionnaire. We just can't determine whether or not they have been enumerated. The Not Classified can be separated into two groups. The first group are the No Contact and Refusals (often referred to as the No Contact, because they are a much larger group than the Refusals) for whom no interview have been completed. Therefore, even though the tracing process has indicated that these persons were enumerable, we don't know their Census day address. The other cases of Not Classified are those for whom an interview have been completed and a Census day address has been obtained but the information we have does not allow us to assign a final classification as either enumerated or missed. It can be because the address corresponds to a Form 4, or the address is too vague. Table 1 gives the distribution of the sample. Only 4% of the SPs receive a final classification of missed, so this is a really rare event that we are trying to measure. In fact, we have the exact same number of persons classified as non-respondent than we have missed. That indicates the importance of having a good non-response adjustment.

Table 1: Final Classification of SPs

Classification	Number of cases
Enumerated	49,198
Missed	2,292
Deceased / Emigrated / Abroad	2,083
Out of scope / overlap	1,151
Not Identifiable	440
Not Traced	1,432
Not Classified	420
No Contact / Refusal	(168)
Others	(252)
Total sample	57,016

It is also of interest to look at the distribution of the non-respondents by frame, because this shows how some parts of the sample are harder to trace and interview. The following table gives the total sample in each frame, along with the number of non-respondents and the non-response rate in each category. The right-most column gives the total non-response.

Table 2: Distribution of Non-Response By Frame

Frame	Sample	NI	NT	NC	Total
Census	42,065	400 (1.0%)	686 (1.6%)	214 (0.5%)	1300 (3.1%)
Births	3,390	13 (0.4%)	97 (2.9%)	13 (0.4%)	123 (3.6%)
Immig.	2,605	0	193 (7.4%)	22 (0.8%)	215 (8.3%)
Missed	2,341	0	125 (5.3%)	32 (1.4%)	157 (6.7%)
NPRs	1,465	27 (1.8%)	224 (15.3%)	37 (2.5%)	288 (19.7%)
HCFs	5,150	0	107 (2.1%)	102 (2.0%)	209 (4.1%)
Total	57,016	440 (0.8%)	1432 (2.5%)	420 (0.7%)	2292 (4.0%)

The overall non-response rate is only 4%, showing how good a job the RRC staff accomplished in treating the cases. It's in the tracing process that the biggest differences emerge, with the NPRs, and on a smaller scale the immigrants, having a much higher rate than the other frames.

5. Adjusting for Non-Response

5.1 User Requirements

It is vital to have a non-response adjustment that is easy to understand and easy to justify. Because the RRC estimates have such a large impact on federal-provincial transfer payments, each step of the study is under high surveillance from users. The adjustment for non-response draws a lot of attention for two reasons. First, the level of non-response is not the same from one province to another, so an adjustment methodology that benefits one province may not benefit another. Second, the uncertainty inherent in a non-response adjustment leaves users with some apprehension about the final estimates.

The following sections will describe the non-response adjustment methodology that was used for the release of the preliminary results in March 1998.

5.2 Basic Weighting Adjustment Groups

The starting point for adjusting for non-response is the formation of Weighting Adjustment Groups (WAG), groups of SPs who have a similar probability of being missed by the Census. We redistribute the total weight of the non-respondents to each respondent in the group in proportion to their share of the total weight of the respondents. This assumes that within a WAG, the respondents represent the non-respondents. A weighting adjustment is done separately for each level of non-response.

Basic WAGs are used to adjust for the Not Identifiable. The composition of these WAGs is easy to determine since there is not a lot of information on the Not Identifiable. The real challenge in the process of adjusting for non-response is to determine what additional information we should use to refine the WAGs to adjust for the other levels of non-response. The basic WAGs are formed by the different values of the replicate (five replicates are used for variance estimation), the frame, the province at selection and the stratum.

In order to avoid distributing the weights of a large number of non-respondents to a small number of respondents, criteria based on a minimum size and a maximum level of non-response have been established. The smaller the WAG, the lower the allowable non-response rate. If these criteria are violated, a new WAG is formed by collapsing the original group with another. The new WAG is created by combining two values of one of the variables used to form the groups, beginning with the stratum. This process is repeated until all the WAGs satisfy the criteria. The criteria are the following:

New WAGs are formed by collapsing with other classes if there is at least one non-respondent in the group and:

$$n<10$$
$$\text{or } 10 \le n \le 12 \text{ and } n_{nr} > 1$$
$$\text{or } 13 \le n \le 15 \text{ and } n_{nr} > 2$$
$$\text{or } 16 \le n \le 17 \text{ and } n_{nr} > 3$$
$$\text{or } 18 \le n \le 19 \text{ and } n_{nr} > 4$$
$$\text{or } n \ge 20 \text{ and } n_{nr} / n > .3$$

where n is the number of persons and n_{nr} is the number of non-respondents in the group.

5.3 Additional Information

More information is available for the Not Traced and the Not Classified persons than we had for the Not Identifiable. We want to take advantage of this extra information in adjusting for the Not Traced and the Not Classified. This section describes the additional information available for these non-respondents.

First, it is known that the Not Traced and Not Classified persons are not deceased since a linkage to mortality files has already identified deceased SPs. Therefore, we can exclude the deceased cases when adjusting for the Not Traced and Not Classified. Also, even though it has not been possible to assign a final classification, we do have some address information. The selection address is known, and, for most of the sample, the update address is known. We have also collected a set of possible addresses from different administrative sources such as telephone directories. For the Not Classified persons (excluding the No Contact and Refusal), we even have some addresses that come from an interview.

For each of the addresses listed above, the Census database was searched to find the Census questionnaire. The questionnaire and associated visitation record were reviewed to determine whether or not the SP was enumerated. If a non-respondent SP was found enumerated at one of the addresses, the SP's status changed from non-respondent to enumerated. For the Not Traced, the No Contact and the Refusals, even though we have looked at all these addresses without finding them, we cannot classify them as missed since an interview is required to indicate where the SP was living on Census day. If an interview is conducted and we then cannot find the SP at their Census day address nor elsewhere, we classify the SP as missed. For the non-respondents for whom we don't have an interview, it is possible that there is an address somewhere where the person is listed on a Census questionnaire, but this address can only be obtained by a direct interview, not through processing administrative data. For the Not Classified that are not No Contact, it is different because they have provided their Census day address, but this information can't lead us to a final classification. The

next two sections give the treatment for the second and third types of non-respondents.

5.4 Adjusting For Not Traced

More is known about the Not Traced, the largest group of non-respondents, than about the Not Identifiable. First, we know that they are not deceased. It is also known that they are not enumerated at any address that could be obtained by administrative means (i.e. any address other that those that could be obtained only from an interview). Let us designate these addresses as "old addresses" and denote addresses that can only be determined from an interview as "new addresses".

Using this information, we create the WAGs for the Not Traced adjustment by taking the same WAGs created for the Not Identifiable and removing each SP that is either deceased or enumerated at an old address. The weights of the Not Traced persons are then redistributed among the remaining respondents in the same WAG.

By doing this adjustment, we assume that, should have we traced these persons, they would have ended up as either missed, enumerated at a new address, emigrated, abroad, out of scope or Not Classified. We then assume that, within a particular WAG, the Not Traced persons will be represented by the respondents falling in one of these categories.

5.5 Adjusting For Not Classified

5.5.1 No Contact and Refusals

The only thing that differentiates the No Contact from the Not Traced is that the tracing process has confirmed that these persons were living in Canada on Census day, i.e. they are enumerable (either enumerated or missed). We create the WAGs for the No Contact in the same way as for the Not Traced, but we exclude SPs that are not enumerable. The weights of the No Contact are therefore distributed among the SPs in the same WAG that are either missed or enumerated at a new address.

5.5.2 Other Not Classified

These non-respondents are very different from the preceding ones because for them, we have obtained a Census day address. This address can either be new or old. Because we know that these Not Classified are enumerable, we keep only the enumerated and missed persons for the weight adjustment, and we separate them in two groups; those for whom the classification address is an old address and those for whom it is a new address. The Not Classified with a new address have their weight redistributed among enumerated and missed respondents with a new address and similarly for the Not Classified with an old address. This is done within the same WAGs created for the adjustment of the Not Identifiable after removing every SP that is not either enumerated or missed.

6. Non-Response Adjustment Results

In this section, we will analyse some of the results of the non-response adjustment that we just described. We can first look at the importance of the WAG collapsing criteria presented in section 5.2. Comparing the number of WAGs before and after collapsing, we see that it is only for the Not Traced that a significant level of collapsing occurs; the number of groups after collapsing is only two thirds of what it was prior:

Not Identifiable	⇨ 99%
Not Traced	⇨ 67%
Not Classified - No Contact	⇨ 93%
Not Classified - others	⇨ 95%

The most important evaluation of the non-response adjustment is to assess the reasonableness of the implicit undercoverage rate associated with each level of non-response. Before we do any non-response adjustment, we have a sampling weight for each unit which is the inverse of the probability of selection (for the missed frame, the final weight of the SP in the 1991 RRC is considered the sampling weight). We can compute an initial undercoverage rate by dividing our estimate of missed persons, calculated using the sampling weight, by our estimate of enumerated and missed persons. Then we redistribute the weight of the Not Identifiable SPs to other people in the sample. A part of this weight is redistributed to enumerated or missed persons. We do the same thing for the other types of non-response. We can compute an implicit undercoverage rate for each non-response type by dividing the part of the weight of these non-respondents that is distributed to missed persons by the part of the weight that is distributed to enumerated or missed persons. Then we can compute a final undercoverage rate by dividing our estimate of missed persons, using the final weights, by the estimate of enumerated and missed persons. The following table gives the implicit undercoverage rates for each non-response type by frame. NC1 is for the No Contact / Refusal and NC2 is for the rest of the Not Classified.

Table 3: Implicit Undercoverage Rate (in %) By Frame for Each Level of Non-response for the Methodology Used for the March 1998 Release

Frame	Initial	NI	NT	NC1	NC2	Final
Census	2.7	2.7	15.6	18.5	5.5	2.9
Births	2.7	2.2	6.2	8.1	8.3	2.8
Immig.	9.2	- - -	22.2	38.2	8.8	10.1
Missed	12.1	- - -	24.5	26.0	14.0	12.9
NPRs	22.6	16.2	50.2	49.2	36.5	25.7
HCFs	5.3	- - -	21.5	23.9	8.0	5.6
Total	3.3	2.8	17.2	20.8	7.6	3.6

The implicit rate for the Not Identifiable is about the same as the initial rate. This was expected, because of the way the adjustment is done for these cases. The only reason why these two rates are not exactly the same is because the initial rate within the WAGs where the Not Identifiable cases are is not necessarily the same as the overall rate for the whole frame.

Relative to the Not Identifiable and the NC2, a lot of the weight of the Not Traced and No Contact is distributed to missed persons. This is especially so for the NPR frame, where we can see that the Not Traced and No Contact are considered having almost a 50% chance of being missed. A reasonable hypothesis is that a person who is harder to trace or contact is also a person that is harder to enumerate on the Census. Moreover, the fact that one of the only significant information that we have about the Not Traced and No Contact is that we verified some of their potential addresses without finding them enumerated there leaves them more chance of being missed.

The lower implicit rate for the second category of Not Classified can also be easily explained. Let us say, for example, that for an SP in the second category of Not Classified persons, the Census day address corresponds to a Form 4 and, therefore, it can't be determined if the SP is enumerated or missed there. Because the SP told us that this was his/her Census day address and a Census form has been filled for this address, the chances that the SP has been enumerated there are higher.
We can see that even though some non-respondents are implicitly given a high probability of being missed, the overall impact on the undercoverage rate is not that big, going from 3.3% to 3.6%. Nonetheless, the number of missed persons added by the non-response adjustment is around 130,000, going approximately from 940,000 initially to 1,070,000 after adjusting for non-response.

7. Modified Adjustment

Other ways of adjusting for non-response have been studied. One of these adjustment have been chosen to be implemented for the release of final estimates of Census coverage error. The purpose of this alternative methodology is to fine-tune the adjustment for the Not Traced persons. In order to do so, we look at additional information (like where do we find the members of the selection household of the SP) to distinguish those Not traced that are for sure at a new address from those that could still be at an old address. This distinction has a big impact on the proportion of the weight of these Not Traced that is reassigned to the missed persons. The following section describe the new methodology for the Not Traced adjustment, with the strengths and weaknesses of this scenario compared to the one presented in section 5.

This scenario uses the same adjustment as the previous one for the Not Identifiable persons and the second type of Not Classified. The logic supporting the proposed changes is the following. It is defensible to exclude the SPs that are enumerated at an old address from the adjustment for the Not Traced persons because we verified these old addresses for the Not Traced persons and haven't found them on the Census questionnaire corresponding to that address. But a problem may arise when we keep all the persons that are missed at an old address, saying that even though we haven't found the Not Traced persons at their old address, they can be missed there. This may be true for some of the Not Traced, but not for all. If, for a Not Traced, we check the Census questionnaire corresponding to an old address and find some members of the selection household of the SP, but not the SP, we can conclude that it is possible that the SP was missed at that address. The same logic holds when the old address has been listed as unoccupied, or hasn't been listed, i.e. the Census enumerator missed it.

On the other side, if for a Not Traced, we check the Census questionnaire corresponding to an old address and we don't find the SP or any member of their selection household, but instead we found totally different persons, the chances are low that the SP was missed at that address. It is more likely that the person has moved and is enumerated or missed somewhere else. The previous methodology was thus implicitly giving these persons too high a chance of being missed. The same logic can be applied to the No Contact persons.

The proposed adjustment would then be the following. For the Not Traced, if one of the old addresses falls into one of these three categories:

1) At least one member of the SP's selection household has been enumerated at that address;
2) The address has been listed as vacant;
3) The address hasn't been listed;

then the adjustment would be the same as the previous one. We will call these not traced the "potential non-movers". The other Not Traced persons would be considered as having moved and their weight would be redistributed to only the persons classified at a new address (Basically the same adjustment as in the previous method, but excluding the persons that are missed at an old address.). The same rule would be applied to the No Contact persons. 24% of the Not traced have been classified as potential non-movers, compared to 32% of the No Contact.

This alternative methodology would result in a lower implicit undercoverage rate for the Not Traced and the

No Contact, thereby reducing the estimate of missed persons. Besides refining the current methodology to consider who other members of the selection household were enumerated at the processed addresses, this scenario allows us to correct one problem of our current adjustment for the Not Traced persons. In reality, a small number of the old addresses for the Not Traced were too vague to be verified. For these cases, it is unfair to exclude the persons enumerated at an old address by saying they are different from these Not Traced. (The exact way to adjust for them would be to distribute their weight to all the enumerated and missed, regardless of whether their classification address is old or new. However, it is not worth adding an extra step to treat so few cases.) The alternative adjustment allows us to consider these persons as belonging to the second category of Not Traced and thus give them a lower implicit undercoverage rate, a more reasonable choice for these non-respondents.

The problem with this scenario is in the way that it treats SPs who were living alone according to selection data. They may or may not still be living alone at the time of Census. Under the proposed alternative scenario, they are considered as all belonging to the second category, unless their old address is of the 2nd or 3rd type listed above in this section. There is room for improving this alternative for SPs living alone at the time of selection.

8. Results for the Modified Adjustment

The following table gives the implicit undercoverage rates for the modified adjustment. The initial rates and the rates for the Not Identifiable are not given because they don't vary from those presented in Table 3 since the previous and modified adjustment are similar for the Not Identifiable. NT1 is for the Not Traced that are potential non-movers while NT2 is for the rest of the Not Traced. NC1 is for the No Contact that are potential non-movers while NC2 is for the rest of the No Contact, plus all the other Not Classified.

Table 4: Implicit Undercoverage Rate (in %)
By Frame for Each Level of Non-response
for the Modified Methodology

Frame	NT1	NT2	NC1	NC2	Final
Census	16.2	8.6	19.4	6.7	2.8
Births	5.8	3.9	7.7	7.6	2.8
Immigrants	19.3	12.4	- - -	12.0	9.5
Missed	22.1	13.3	24.2	14.0	12.3
NPRs	53.0	30.5	27.3	32.8	23.5
HCFs	24.3	10.2	23.1	9.2	5.5
TOTAL	16.6	9.7	19.1	8.9	3.5

The rates in the column NT1 are similar to those of the column NT of Table 3. Sometimes they are (slightly) higher, sometimes lower. It is because the adjustment for NT1 (the Not Traced that are potential non-movers) is the same as what it was for the Not Traced in the previous adjustment. The fluctuations come from the fact that in the new adjustment, we have removed some Not Traced to place them in the other category (NT2). These Not Traced were sometimes in groups with higher rate, sometimes with lower rate. The same observations can be made when comparing the column NC1 of the two tables. The only surprise in this case is for the NPRs, where the rate goes down by almost 50%. This is caused by the fact that the modified adjustment leaves only two SPs in that category, making the rate very unstable.

The numbers in the NT2 columns are significantly lower than those in the NT1 column. This was expected because the adjustment is the same for these to types of Not Traced, except that for the second type, we remove some missed (those missed at an old address) from the adjustment, thus decreasing the implicit rates.

The numbers in the NC2 column are similar to those same numbers in Table 3, only slightly higher in most cases. This is because this adjustment is done in the same way in both cases. The only difference is that for the new adjustment, the No Contact that are in the second category (those that are not considered as potential non-movers) are adjusted at that step, along with the second type of Not Classified. In the previous adjustment, only the latest where adjusted at that step. Because these No Contact are all adjusted with the enumerated and missed at a new address, for which the undercoverage rate is in general higher than for those enumerated and missed at an old address (persons more mobile are more likely to be missed), it cause the implicit rate to be in general slightly higher for NC2 in this new adjustment than it was in the previous one. Overall, this modified methodology brings down our estimate of missed person by about 40,000 persons.

9. Conclusion

Adjusting for non-response in the Reverse Record Check Study is a delicate and complex operation. Short of having a separate adjustment for each non-respondent, it is challenging to take into account every piece of information that could lead to the best adjustment possible. The current 5-stage approach gives solid results even though some improvements could be made by trying to get more information on the non-respondents.

For the next RRC, improvements in the collection methodology like the introduction of CATI, better variance estimation, and other changes may allow a better identification of the characteristics of the non-respondents, thereby improving the quality of the non-response adjustment.