Census Coverage Studies in Canada: A History with Emphasis on the 2011 Census

David Dolson
Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6

Abstract

For Canada's 2011 Census of Population, Statistics Canada will again use the Reverse Record Check methodology for estimating undercoverage. Estimates are based upon classification of a sample of persons who "should" be enumerated as being in-scope enumerated, in-scope missed or out-of-scope. Overcoverage will be estimated using the Census Overcoverage Study, first implemented in 2006. This methodology is based upon matching the census database to itself using administrative records to aid in confirming the validity of matching pairs of enumerations.

This paper will trace the history of census coverage studies in Canada, primarily focusing on the period since 1961 when the Reverse Record Check was first implemented. Emphasis will be placed on plans for 2011, outlining the methodology of each study and highlighting important changes from the 2006 coverage studies.

Keywords: Reverse Record Check, Record Linkage

1. Introduction

The first census in Canada including a formal coverage measurement program was that in 1961; undercoverage was estimated using the Reverse Record Check (RRC) methodology. This paper will first give a description of coverage measurement fundamentals and concepts as applied in Canada. The next section will review the methodology of the 1961 RRC and trace major developments in the coverage studies program over the following 45 years. The last section will provide a detailed description of plans for 2011 including the Dwelling Classification Survey, the RRC and the Census Overcoverage Study. Use of results in the Population Estimates Program is briefly noted.

2. Fundamentals and Concepts

The target population for the Canadian Census of Population includes persons with a usual place of residence in Canada who are Canadians, landed immigrants or non-permanent residents (refugee claimants, persons with a work or study permit covering Census day) as well as two small groups of persons not currently living in Canada. A modified *de jure* method of enumeration is used in which a key modification is that students and workers who spend most of the year elsewhere but who return periodically to live at home are to be enumerated at that home.

Let T be the size of the **target population**, C be the published **census count** and N be the **net population coverage error**. Then $T = C + N$.

Let U be **population undercoverage,** the number of persons not in C who should have been and O be **population overcoverage,** the number of enumerations in C that should

not have been there. It consists of out of scope and duplicate enumerations of in-scope persons. The first component is believed to be small in Canada and is assumed to be zero for the coverage studies; measurement of overcoverage focuses on the latter. So $N = U - O$. The objective of the coverage studies is to produce good quality estimates, $\hat{U}$ and $\hat{O}$, at various levels of aggregation. And we get $\hat{T} = C + \hat{N} = C + \hat{U} - \hat{O}$.

$\hat{N}$ is also an important output of the coverage studies program and has been an input to the population estimates program since 1991. For estimating coverage error at sub-national levels, the coverage studies view the census usual place of residence rules as guidelines and persons who are enumerated exactly once are deemed to have been enumerated in the correct place, thus not affecting coverage estimates.

The Census count C is composed of two components: $C = E + I$ where E is the number of **enumerations** and I is the number of **imputed persons**.

The Dwelling Classification Survey produces estimates of persons in dwellings classified as occupied that did not respond to the Census and of persons in occupied dwellings mistakenly classified as not occupied. On the basis of these data, persons are imputed onto the Census database: into non-responding dwellings and into an appropriate proportion of dwellings classified as not occupied. I is the total of these imputations. Note that while I accounts for persons who were in fact not enumerated, it is considered to be part of the Census count and not a component of coverage error.

This has ramifications for the RRC. Let M be the number of in-scope persons who are **not enumerated**; thus $T = E + M$. Its estimate $\hat{M}$ is computed directly from the RRC: any sampled person who is in-scope and not found enumerated (i.e. in E) is classified as not enumerated and so $\hat{M} = \hat{U} + I$. Thus the number of imputations must be netted out of the estimate of not enumerated to arrive at the estimate of population undercoverage.

3. The History

3.1 1961-1986

While the Reverse Record Check might presently be widely viewed as "the Canadian approach to census undercoverge estimation", initial development of the methodology took place at the United States Census Bureau (USCB) in the 1950s. The USCB had used a post-enumeration survey (PES) for evaluation of coverage error in the 1950 Census and in the late 1950s was making its plans for coverage assessment for 1960. This included both further development of the PES for 1960 as well as consideration of interpenetrating samples and related ideas that would lead to the development of the RRC methodology. After a visit to the USCB in 1959 to discuss coverage measurement methodology Ivan Fellegi, future head of Statistics Canada, recommended the RRC methodology as the more appropriate for use with Canada's census. One of the major reasons was that Canada's quinquennial census made the RRC methodology – and its use of the previous census as a frame – much more viable then with a decennial census. Another reason was concern about effectiveness of the PES in 1950 (National Research Council, 2008). U.S. national estimates of net undercount for the 1950 Census derived via demographic analysis (3.5%) and via the PES (1.4%) differed widely from each other. Despite

substantial care taken in data collection operations to ensure independence of the PES enumeration, this difference was attributed to correlation bias in the PES.

The coverage check for the 1961 Census is documented in Fellegi (1968). The only study was the RRC and the objective was to estimate the under-enumeration in the 1961 Census of Population for the purposes of evaluation of the 1961 Census and for providing recommendations for the 1966 Census.

In the RRC a sample of persons from frames covering the census target population is surveyed and processed against census returns to determine if the sampled persons were enumerated, not enumerated or out of scope. An estimate of underenumeration can then be derived. The only frame used for 1961 was persons enumerated in the 1956 Census and the study was limited to the 10 provinces. Due to practical difficulties frames for intercensal births and immigrants were not readily available and no frame existed for persons missed in the 1956 Census. A self-weighting three stage sample of 6,237 persons was selected within 127 enumeration areas within 27 federal electoral districts. A clustered design was used to: reduce clerical effort in sample selection and clerical transcription of information from 1956 Census returns as well as in searching 1961 Census records and to reduce expected travel expenses during RRC field activities.

The first task was to obtain the address for each selected person (SP) in 1961. This involved several steps, starting with mailing of a registered letter to each SP at the 1956 address. There was no questionnaire; the SP was simply asked to reply, indicating their current address. After a follow-up letter if necessary, post office returns were sent a letter addressed to the householder seeking a new address for the SP. If one was provided, a letter was sent. About two thirds of the sample was successfully traced by these steps. Remaining cases of non-response or response with no new information were sent to regional offices for further very intensive tracing, including personal interviewing; about 85% of these cases were successfully traced.

The next step was an intensive search the 1961 Census records – a large clerical operation – to classify each SP as enumerated, not enumerated or out of scope (e.g. deceased, emigrated). In the end all but three percent - "tracing failed" – were successfully classified. Over subsequent censuses, this figure became an important feature in evaluating the success of RRC data collection and SP classification; an extremely high response (classification) rate is required for adequate quality in the RRC estimates.

Direct estimation was used with the weight of non-respondents prorated across all respondents. The estimated missed rate for the 1961 Census was 3.3%.

There were two important observations that became fleshed out in more depth in later years and which led eventually to improvements to the methodology. First, it was noted that the proportion of not enumerated increased with the difficulty of tracing the SP – an observation supporting the very intensive effort in data collection and classification. A related point noted by Fellegi was that the tracing failed group was very likely a group with special attributes and of particular interest.

Based upon the success of the 1961 RRC, the study was significantly expanded for 1966 (Muirhead, 1969). A substantially increased sample of 26,500 was selected from four frames: persons enumerated in 1961, intercensal births, intercensal immigrants to Canada

and persons missed in the 1961 Census as represented by persons - with their final weights - detected as missed by the 1961 RRC.

In the Census frame, a stratified two stage sample of persons within enumeration areas (EA) was selected. The design was self weighting at the provincial level. Within each sampled EA a systematic sample of persons sorted by age and sex was selected; so within each province, the sampling probabilities were the same for all ages and both sexes. The birth and immigrant frames were stratified by province and year of birth or immigration respectively and the same sampling fractions were used as in the Census frame. The entire missed frame was included in the sample.

Tracing, searching, classification of SPs and estimation were all conducted in much the same way as 1961. Mason and Ashraf (1969) gives a detailed explanation. Non-response adjustment groups were defined by frame, province, sex and age group. Weights were also adjusted so that RRC estimates equaled population totals, where known, at the level of the non-response adjustment groups or higher. In 1966 the tracing failed rate increased to 3.4%. Nonetheless and as expected the estimated missed rate of 2.6% was measured with much improved precision.

For 1971 some important improvements, both statistical and operational, were implemented for the RRC (Gosselin, 1976; Brackstone and Gosselin, 1973). With only a small increase in sample size, survey objectives for population undercoverage remained unchanged. A new objective was to produce estimates of household undercoverage as well. The same four frames were used.

Reflecting increasing understanding of coverage error and the availability of computers to facilitate use of more sophisticated designs, the survey design underwent some improvements. The census sample was again selected in two stages, this time in replicates at the first stage. Stratification was unchanged for births and immigrants. Instead of a uniform sampling rate for all age-sex groups, a two fold higher rate (1/500) was used for demographic groups that had higher missed rates in 1966 – immigrants, persons aged 20-24 in 1971 and babies less than 1 year old. All others were sampled at the rate of 1/1,000.

In an important cost saving initiative, field data collection was no longer done for the entire sample. Instead, in a first step, all SPs were matched at their 1966 address to 1971 census records as part of regional office census processing. This step found about 42% of the sample enumerated and only the remaining 58% were sent to regional offices for tracing and data collection. Otherwise processing and classification were generally similar to 1966 with use of registered letters etc. although there was an important increase in the use of administrative data to facilitate tracing. For example, sampled birth registrations were traced forward using family allowance admin data and sampled immigrants were traced using admin records on Social Insurance Numbers. Checks against Old Age Security records and death registrations were also used. For a first time during regional office tracing, telephone tracing was undertaken prior to any personal interviewing attempts; it was very effective.

Analysis and evaluation of the 1971 results produced some useful observations and a number of specific recommendations for 1976. The main observation was that persons with high missed rates tended to be groups with high rates of mobility, e.g. persons not related to the head of the household, males age 20-24, recent immigrants, unemployed persons and small households in rented dwellings. Key recommendations included:

increase the sample size to 33,000 so as to facilitate designing for reliable estimates for each province and certain large subgroups and to allow better analysis of the characteristics of missed persons; expand the use of admin data in tracing to include tax records; replace the use of registered mail for tracing by an expanded telephone operation; implement a new coverage study to produce estimates of misclassification of occupied dwellings as vacant.

The recommendations from 1971 were all implemented. Statistics Canada (1980) provides an overview of the 1976 coverage studies program and its results; a similar detailed technical report would be published following each subsequent round of census coverage studies.

The sample design of the 1976 RRC was similar to 1971 with some small improvements to the stratification in the census frame. The recommended improvements to data sources, tracing methods and data collection were implemented; following tracing, SPs were interviewed by telephone or in person using a questionnaire to collect addresses and some characteristic information.

The principles and methods for estimation were not changed significantly; Gosselin and Théroux (1979) provides a detailed report on the estimation methodology for the 1976 RRC. Despite the modernized methods and sources for tracing the tracing failed rate deteriorated to 4.8%.

The Vacancy Check Survey (VCS) was introduced in 1976 to estimate the number of occupied dwellings misclassified as vacant and to estimate the number of persons not counted in the census as a result. A stratified two stage sample of about 1,400 enumeration areas was selected and all dwellings that had been classified as vacant by census enumerators were revisited about a month after Census day. By interviewing a knowledgeable occupant or neighbour, the VCS interviewer was to determine if on Census day the dwelling was occupied or vacant. This survey was predicated on the assumption that the VCS interviewer focusing on only this task would do a better job than the census enumerators for whom this classification had been one amongst many other responsibilities. Survey estimates were that 7.5% of "vacant" dwellings were actually occupied on Census day. No adjustments were made to the 1976 census figures on the basis of these results. Starting in 1981, a weighting procedure was used to adjust census figures to account for these misclassifications.

For 1981 and 1986 both the RRC and the VCS were repeated with essentially unchanged methodology; see Burgess (1987), Carter (1988) and Statistics Canada (1990).

For the RRC however, there was a major improvement implemented in the non-response adjustment for both these years where, for the first time, the "special attributes and particular interest" of tracing failed cases was recognized in the estimation procedures. Based upon evaluations from 1976 and earlier, it was assumed for 1981 that all persons in the sample from the 1976 census frame who had not moved from their 1976 address and who were enumerated in 1981 would be successfully traced and identified as enumerated by the RRC. And so the weight for the tracing failed group was allocated to all respondents except that group. An important improvement operationally was that the tracing failed rates in the 1981 and 1986 RRCs were 3.4% and 3.8% respectively, significantly reduced from 1976.

In 1986 an improvement to tracing effectiveness was that for the first time SPs from the 1981 census frame were traced forward to a more current address using income tax records prior to initiating the regional office process of matching to 1986 census records to identify enumerated SPs.

Burgess (1987) made a number of other comments and recommendations that are worth repeating here.

Since they are not included in any of the frames, RRC estimates of missed are biased downwards to the extent that persons in the following three groups are missed in the census: illegal aliens; Canadians who were out of scope at the time of the last census (e.g. emigrated) but who had since returned to Canada; and persons born prior to 1966 who had never been enumerated in a 1961 or later census. These populations were estimated at >50,000, about 85,000 and "very close to zero", respectively.

Consistent with an earlier conjecture, Burgess (1987) gave estimates for 1981 that demonstrate that the missed rate for long distance movers is higher than that for local movers which in turn is higher than that for non-movers.

Recommendations included the following: conduct an experiment in 1986 to estimate overcoverage; initiate research into implications and methods for incorporating estimates of net coverage error into population estimates; further improvements to RRC tracing by tracing both SPs and their household members as a means of locating the SP and by initiating admin file tracing prior to Census day.

Carter (1988) clearly hinted at a potential future move to adjusting population estimates for net undercoverage. He discussed not only overcoverage measurement but also means – such as the error of closure - for analysis of coverage studies estimates of net undercoverage viz-à-viz estimates from demographic analysis. Further, he noted potential distortions in estimates and analyses of population growth when undercoverage changes between censuses or is differential by demographic or geographic group. He asked "Should the Census results not then be adjusted for undercoverage"? But he also noted that adjustment could result in worse quality counts and distributions.

This issue was considered in depth for the following several years, ultimately resulting in the decision that with the release in 1993 of estimates of coverage error for the 1991 Census to start adjusting population estimates for net undercoverage while still leaving the Census database itself unadjusted for coverage error. Royce (1992) and Dick (1998) describe statistical methods for assessing the questions raised by Carter.

3.2 1991

1991 was an important census for the coverage studies (Statistics Canada, 1994) and the population estimates program. In addition to the above major decision, non-permanent residents (e.g. foreign students and workers) became in-scope for the census and thus also for the coverage studies. A program to estimate overcoverage became necessary so that estimates of net undercoverage could be produced. The objective of the coverage studies was changed to become the production of good quality estimates of undercoverage and overcoverage for Canada, each province and territory, and important subgroups. The population estimates program had to develop the methods to incorporate estimates of coverage error; this is briefly outlined in section 5.

It was also the start of the current era where there is a close collaboration and consultation with representatives of provincial/territorial statistical offices regarding coverage studies methods and evaluation of coverage studies results.

The RRC had about a 60% increase in sample size to about 56,000 so as to provide estimates of adequate quality for use in adjustment as well as to include the two territories (it later became three in 1998 with the creation of Nunavut) and the non-permanent residents in the RRC. For the former, health care files maintained by the two territories were used as frames. Due the high migration rates for the territories it was felt these files would provide better results than using the set of frames used for the provinces. Non-permanent residents all require permits issued by immigration authorities and so a frame of persons with a permit including Census day was readily available.

The RRC design was updated to address the new objectives. Two thirds of the sample was allocated to provinces and territories to achieve equal precision estimates of undercoverage. The remaining third was allocated using Neyman allocation, addressing the national objective. Both these allocations accounted for estimated population sizes and expected undercoverage rates. Within provinces/territories and frames the sample design was largely the same as in past cycles. Remaining steps of tracing, searching the census database, classification and estimation remained much the same as in 1986.

Three studies were implemented to estimate overcoverage. The Private Dwelling Study was a survey in which a sample of persons enumerated in 1991 was contacted to collect additional addresses where they may have been enumerated. This study was used to estimate overcoverage in one person households or when enumerations were not in the same enumeration area. The Automated Match Study (AMS) was implemented to measure overcoverage within EAs between households of at least two persons. Within each of a sample of about 9,500 EAs pairs of households with similar characteristics by sex and date of birth of household members were identified by computer matching of 1991 census records. This “MonsterMatch” methodology for matching households and scoring the quality of those matches became a valuable standard tool for the RRC as well starting in 1996. Based upon the quantity and match quality of such pairs the EAs were stratified by the probability of including overcoverage. Then for a sub-sample of the EAs the census records for the household pairs were clerically verified for the presence of overcoverage. Finally, the Collective Dwelling Study was implemented to measure overcoverage of persons between collective dwellings and private dwellings. Alternate addresses were collected for a two stage sample of persons enumerated in collectives. Census records for the alternate addresses were verified for overcoverage. All three studies used direct estimation methods to arrive at an estimate of 0.56% overcoverage.

The 1991 VCS and the related methodology for adjusting the census database were unchanged from previous cycles.

3.3 1996

Several major changes were made to the RRC. Taking advantage of improved computing power the scope of the AMS was expanded considerably. The Collective Dwelling Study and the VCS both remained unchanged. It is all outlined in Statistics Canada (1999).

For the first time, the RRC sample for the census frame was selected as a one stage sample of persons. Sample allocation to provinces/territories remained as in 1991. The sample for the missed frame is determined by the results of the 1991 RRC. Otherwise though, below this level the frames were treated all together and the design was based on demographic stratification into groups with similar undercoverage rates and optimal allocation, accounting for both historical trace rates and expected undercoverage rates. The result of course was a much more efficient design than previously.

1996 was the start of an extended period of intense focus on non-response – the tracing failed cases – adjustment in the RRC. It was increasingly believed that the approach used up to 1991 was too simple and that improvement was needed to reduce associated biases. To that end a more refined procedure was implemented first by partitioning the tracing failed group into three sets: not identified SPs with such poor identification information that tracing and classification were impossible; not traced SPs who could not be contacted and interviewed; and not classified SPs who were interviewed and detemined to be in scope but who could not be classified because address information was too vague. A three step adjustment procedure was implemented where first, within non-response adjustment groups, the weight of the not identified was allocated to all other SPs. Then the updated weight of the not traced was allocated to all respondents. Finally the updated weight of the not classified was allocated to in scope respondents. In these last two steps, SPs classified enumerated who had not recently moved were excluded from the non-response adjustment. Further improvements were made in each of 2001 and 2006.

The Private Dwelling Study was incorporated into the RRC. Since the RRC was now also used to measure overcoverage, regional office processing of the RRC sample against census returns was cancelled and all SPs were to be interviewed to collect household information and a variety of addresses where the SP could be enumerated. To minimize the need for tracing by interviewing staff, addresses from the frames were updated via linkage to admin files prior to telephone interviewing. Following interviewing, census questionnaires were searched for all enumerations of each SP using household information from both the frame and the RRC interview as well as all available addresses of the SP (and members of the household from the frame). Although this remained a laborious clerical operation, for the first time it was assisted with some automated matching of 1991 households of SPs in the RRC sample to households enumerated in the 1996 Census using the MonsterMatch (Mayda, 1998) available from the 1991 AMS.

In an expansion of an already very thorough operation, a RRC follow-up survey was implemented where each SP who, after a first effort at searching the 1996 census database, appeared to be in scope but not enumerated was again interviewed to seek additional address information. It proved to be worthwhile. The new information thus collected facilitated the classification as enumerated of many SPs who would otherwise have been incorrectly classified as not enumerated.

Since the AMS could be much more precise than the RRC for overcoverage it was decided to take advantage of improved computing power to considerably expand the study while reducing dependence on the RRC. For 1996 it was converted to a one stage sample design instead of two and was expanded to measure overcoverage between households of at least 2 persons within regions of one or more provinces (Ha et al, 1998). Within each region the Census database of households was matched to itself using the MonsterMatch. Pairs of households were stratified by province, geographic proximity, household size and quality of match. Direct estimates were produced based on clerical

verification of a sample of 7,700 pairs of households. The increase in the estimate of overcoverage to 0.73% was attributed entirely to improvement in its estimation.

3.4 2001

For a second cycle in a row there were several major changes, many of them related to improved automation. Again the Technical Report on Coverage (Statistics Canada, 2004) documents this.

Except for some fine tuning of the stratification and sample allocation the design of the 2001 RRC was the same as in 1996. The sample size was increased to about 61,000 and the sample allocation was modified slightly to account for the fact the RRC was being used to estimate overcoverage as well as undercoverage. A substantially improved non-response adjustment methodology (Théberge and Liu, 2002) was implemented.

The non-response concepts of not identified, not traced and not classified were retained. The 1996 procedure had addressed the issue of the ease of tracing an SP by whether the SP had moved recently. The 2001 procedure recognized that it was not actually important whether the SP had moved. Instead the non-response adjustment model was based, in part, on the observation that when the Census day address is known prior to interviewing the SP is easier to trace and more likely to be enumerated while when it is not known prior to interviewing the SP is harder to trace and less likely to be enumerated. A new parameter was also introduced to the model to recognize that SPs whose census day address was known to the RRC prior to data collection should be interviewed at about the same rate whether enumerated or missed. This parameter was estimated by the response rate in data collection for SPs who had been classified as enumerated at an address known prior to data collection. Consistent with previous observations by Fellegi regarding how special the not traced group is, this new model thus included an adjustment to account for differing response probabilities depending on whether the SP would be: enumerated or missed at a location we knew without interviewing; enumerated or missed at a location we had to determine via interviewing; or out of scope (other than dead). This did a superior job of ensuring the adjustment procedures were as unbiased as possible.

Much increased use of technology in operations helped reduce non-sampling error in the RRC. For 2001 RRC data collection was done using CATI for the first time. As well, RRC interviewers were able to take advantage of automated tools to assist in tracing that had been developed for use by Statistics Canada's longitudinal surveys. Automation had an even greater impact for RRC processing where clerical staff no longer had to search through paper documents. Instead RRC clerks sitting at work stations had electronic access to RRC data, images of census questionnaires, mapping software and assorted other tools to assist in searching and classification of SPs. This facilitated a very significant improvement in processing quality and efficiency.

Overcoverage was again measured via the same three studies as in 1996. The sample size of pairs in the AMS was more than doubled to about 17,000.

Although its methodology remained largely unchanged the Vacancy Check Survey was renamed as the Dwelling Classification Survey and its scope was expanded to address all types of error in classification of dwellings. Its previous objectives were retained and addressed by the same procedures: estimation of dwellings classified as unoccupied that were in fact occupied on Census day and to provide adjustments for census data to

account for these errors. New objectives were: to estimate the number of census non-response dwellings that were actually unoccupied on Census day; to estimate the number of persons living in non-response dwellings; and to adjust the household size distribution on the census through whole household imputation for the non-response dwellings.

3.5 2006

For the first time in Canada, the 2006 Census data captured names from census questionnaires and they were available during processing on a response database. This facilitated a number of improvements for the coverage studies (Statistics Canada, 2010).

For 2006 the RRC returned to its roots and was used only for measuring of undercoverage. This made processing simpler than it had been in 1996 and 2001; once a SP was found enumerated, no further searching of the Census database was necessary. It made data collection cheaper since the RRC could return to the approach of only interviewing those SPs who could not easily be found enumerated on the census database. However, to meet the needs of the non-response adjustment model it was necessary to also send for RRC interviewing a subsample of SPs classified enumerated at an address known previously.

For the 10 provinces the RRC sample design was modified only slightly. The sample size was increased somewhat to 68,000 and its allocation to provinces was oriented further towards equal precision of provincial estimates than had been the case for 1991-2001 with 85% of the sample allocated with that objective and only 15% allocated to the national optimization. Optimal allocation was again used within provinces but an improvement was made by using a raking procedure developed by Théberge (2006) to better stabilize some design parameters, such as the expected missed rate by stratum, and make them consistent with values used at the provincial level.

In each of the territories a new design – facilitated by the census data capture of names - was used. The entire frame derived from each territorial health care file was matched to the Census response database by name, age and sex using exact matching with clerical verification. Matches (i.e. enumerations) were each given a weight of one and a stratified sample of non-matches was selected for interviewing and further processing to classify them as enumerated, missed or out of scope. Precision of estimates from the territories was much improved with this new design.

CATI was used for data collection. New for 2006, mailout and personal interview collection were used in special circumstances where an SP could be definitively traced to a specific location but a telephone response could not be obtained.

The data capture of names by Census facilitated an important improvement for RRC processing by providing a powerful new variable that could be used in automated record linkage to search for SPs on the Census database using name, date of birth and sex as matching variables. Along with searching at addresses (for both the SP and household members) from the frame, administrative sources and the RRC interview as well as the continued use of the MonsterMatch, a large proportion of enumerated SPs could be reliably classified as such without need for clerical review. These powerful approaches in combination with other improvements in searching capability in 2001 and 2006 meant that a definite determination of an SP's enumeration or not was possible for virtually any address, even a very vague one.

This in turn had the valuable impact that some of the concepts (e.g. not classified) used in the 2001 classification procedures and non-response adjustment model were no longer needed and so a simpler methodology could be used (Théberge, 2008). For a detailed description of concepts and classification procedures see Diallo (2008). For 2006, it remained critical to establish whether information from the RRC interview was needed for determining if the SP was enumerated or not and whether the RRC had the address of classification prior to data collection or not.

The previous approaches for estimating overcoverage were dropped and an entirely new methodology called the Census Overcoverage Study (COS) was implemented (Morel and Farr, 2007) and produced estimates with much improved precision. In a two step procedure using first exact matching methods and then probabilistic methods (Fellegi-Sunter method) the census database of enumerations was matched to itself with the assistance of a file assembled from administrative sources covering much of the population. The AMS was retained for evaluation purposes.

For 2006, the design and estimation procedures for the DCS remained unchanged. However, the procedure for making adjustments for the census was simplified by expanding the use of whole household imputation to account for occupied dwellings misclassified as vacant in addition to its use to account for non-responding dwellings.

4. Plans for 2011

Coverage studies for 2011 will include the Dwelling Classification Survey, the Reverse Record Check and the Census Overcoverage Study.

The DCS will be largely unchanged from 2006. It will provide data for use in adjusting the census database for non-responding dwellings and for occupied dwellings mistakenly classified as unoccupied by means of whole household imputation. Guided by the distribution of household sizes and occupied rates observed in the DCS, a form of nearest neighbour imputation will be used to impute whole households to appropriate percentages of non-responding dwellings and dwellings classified as unoccupied on the census database. The survey will be again be taken as a stratified (by province and large urban area versus not) two stage sample of dwellings within enumeration areas. Direct estimation is used. New for 2011, we expect to increase the sample of EAs from the former level of about 1,400 while introducing sampling at the second stage so as to reduce the impact of intracluster correlation on the quality of DCS estimates.

The RRC will estimate the number of persons not enumerated and by subtracting out the number of DCS based imputations, the estimate of persons missed in the Census will be derived. The objective remains unchanged: to produce good quality estimates of undercoverage for Canada, each province and territory and important subgroups.

The RRC will use a set of five frames for the provinces: persons enumerated in 2006, persons classified as not enumerated in the 2006 Census by the 2006 RRC, lists of intercensal immigrants and of NPRs provided by Citizenship and Immigration Canada and lists of intercensal births coming from provincial vital statistics registries. Because birth registration data for 2011 will not be available in time for the RRC and that for 2010 may not be, the RRC is considering using newly available data from federal Child Tax

Benefit files as the frame for persons born in 2010 and 2011. Frames for the three territories will again be extracted from territorial health care files.

The sample design for provinces will be very similar to that in 2006 – a stratified SRS of persons. The 2006 Census frame will be stratified by province, sex, marital status and age group (differing age groupings by marital status). Other frames are stratified by province only. The sample of about 70,000 is allocated to provinces primarily for equal precision of estimates of the missed rate. Within provinces, Théberge's (2006) procedure for optimal allocation with smoothing of design parameters will again be used. The immigrant and NPR frames can overlap with each other as well as with the missed and 2006 Census frames. Using record linkage, steps are taken to remove this overlap.

For the territories, the first step is an exact match by name, sex, and date of birth of the entire frame of about 100,000 persons to the 2011 Census enumerations. All matches – about 80,000 are expected – will be clerically verified. A stratified SRS of about 2,000 of the non-matches will be selected for interviewing and further processing.

For the first time, weights for the sample from the previous (i.e. 2006) Census frame will be adjusted to account for overcoverage in the frame. A potential relative bias of just less than 1.6% is thus removed.

Following sampling for the provinces, frame data for each sampled person and their household members are matched to income tax data and other administrative sources such as driver's license files and an electronic telephone directory (Infodirect) to obtain more recent addresses (and telephone numbers) where the SP might be enumerated.

Once the final Census database is available in October 2011, RRC processing (i.e. searching of the 2011 Census database) is started. This is a large record linkage operation with automated and computer assisted clerical steps in which the goal is to: classify each SP as enumerated (and thus assumed in scope), in scope and not enumerated, out of scope (e.g. dead or emigrated) or not traced (not classifiable as in or out of scope); determine the usual place of residence for persons classified not enumerated; and derive other variables for non-response adjustment.

For each SP the Census database is searched at potentially several addresses identified using: record linkage using name, date of birth and sex; addresses having similar household composition to that on the frame (MonsterMatch procedure); the address from the frame; addresses from administrative sources. First steps are automated and it is expected that about 36,000 SPs will be classified enumerated without need for clerical review. Most deceased persons will be identified via matching to death registrations. Computer assisted clerical processing is then initiated for SPs not classified enumerated. Clerks will have mapping and automated search tools that will help them resolve incomplete addresses and addresses from lower quality record linkage suggestions.

RRC data collection will start in January 2012, again using primarily CATI. All the case information for each SP not classified enumerated in initial processing, including the various addresses noted above, is forwarded to a regional office where further tracing using local resources is done if necessary. This is expected to be about 10,000 cases. An additional subsample of 5,000 to 10,000 SPs already classified as enumerated will be interviewed to facilitate estimation of a parameter critical for the non-response adjustment methodology. The RRC interview collects the census day roster, demographic

data, the census day address and other addresses where the SP might be enumerated as well as information needed to determine if the SP is in scope or not. Interview data are used by processing staff to complete the classification work for each SP.

The non-response adjustment procedure developed by Théberge (2008) will again be used with direct estimation, treating the design as a two phase sample.

The Census Overcoverage Study will again use the methodology implemented for 2006. In a first step all enumerations on the Census database will be exact matched by name, date of birth and sex to a file assembled from tax, birth, territorial health care files and immigration records covering a large proportion of the population. One-to-one matches are considered correct enumerations and many-to-one overcoverage. Samples of both of these will be selected for clerical validation; error rates will be estimated and results incorporated into the estimation procedures. Many-to-many cases are sampled for clerical determination if there is overcoverage or simply correct enumeration of the "many".

In the second step, the non-matches from the Census will be matched back to the full Census database of enumerations using name, date of birth, sex and geographic proximity as the linkage variables. Fellegi-Sunter probabilistic record linkage will be used with upper and lower boundaries for match scores set conservatively so as to minimize the risk of erroneous classification of matches as overcoverage or not. Matches with scores between the two boundaries will be sampled for clerical processing.

Several refinements will be implemented. Samples of matches in step 1 for clerical verification will be larger than in 2006. This is to provide improved precision of estimated error rates that, in 2006, were a bit larger than expected and differential by province. For step two, it is solely to improve precision. If possible, the sample design for step two will be improved by implementing upper/lower boundaries that may differ by province and by some changes to the stratification and sample allocation. There are a number of other minor modifications not described here.

While quality management is always important, it is particularly so for census coverage studies, given their use and the degree of scrutiny. I highlight here some practices that go beyond those one would normally implement for survey data collection and processing. Because successful tracing and accurate classification depends critically on the quality of address and other data a substantial amount of effort is put into cleaning of these data during preprocessing. Successful tracing of difficult cases is very important to the success of the RRC; in this regard, the RRC leverages its success not only on its own experience but also on the lessons learned and the tools developed in the context of Statistics Canada's longitudinal surveys. Substantial attention is paid to the quality of classification decision making by clerical staff. Clerks with experience from Census processing will be hired and further detailed training will be provided on coverage concepts and classification procedures. A three step escalation procedure from clerks to supervisors to very experienced coverage experts will be implemented for difficult to classify cases. A final step in processing is that all SPs classified as not enumerated are subjected to a detailed review and confirmation process by the senior coverage experts.

Following a period of detailed internal review and certification, estimates of coverage error will first be released in March 2013. This is followed by a three month period of intensive collaborative review with representatives of provincial/territorial statistical offices leading to final estimates which will be released in September 2013.

This review and evaluation, done collaboratively by staff from the coverage studies and the population estimates program, focuses on several key indicators of quality; some are noted here. Estimates (and evaluations) are produced for numerous domains at a number of levels of geographic aggregation. Estimates of missed are analyzed for internal consistency and relative to historical patterns while taking into consideration feedback from Census collection operations. RRC estimates of deceased are compared to counts from death registrations. RRC estimates of persons enumerated are compared to the actual number of enumerations in the Census. A particularly important evaluation is the comparison of census counts adjusted for net coverage error to population projections from the previous census (error of closure).

5. Population Estimates Program

Since 1991, Statistics Canada's population estimates program (PEP) has incorporated an adjustment for net undercoverage in the census. To do so the PEP requires estimates of net undercoverage by single year of age and sex for Canada and for each province and territory. The methodology to derive this from the coverage study results as well as the PEP estimation methodology itself are given in Statistics Canada (2007).

Acknowledgement

Thank you to I.P. Fellegi for a very informative discussion on the early days of census coverage studies in Canada and to C. Thibault for comments that improved the paper.

References

Brackstone, G.J. and Gosselin J.-F. (1973). 1971 Census Evaluation Program – Reverse Record Check – Methodology Report. Statistics Canada internal report.

Burgess, R. (1987). Limitations of Reverse Record Check Estimates of Census Undercoverage. Statistics Canada internal working paper.

Carter, R.G. (1988). Measuring Coverage Errors in the Census of Population. Statistics Canada.

Diallo, M.S. (2008). Contre-verification des dossiers Recensement 2006: Classification 2006 et fichier maitre des personnes choisies 2006. Internal report. Statistics Canada.

Dick, J.P. (1998). Testing Strategy using the 1996 Coverage Study Results. Statistics Canada internal working paper.

Fellegi, I.P. (1968). Coverage Check of the 1961 Census of Population. Technical Memorandum (Census Evaluation Series) No.2, Dominion Bureau of Statistics.

Gosselin, J.-F. (1976). The Methodology of the 1971 Reverse Record Check. *Survey Methodology*, 2, 180-194.

Gosselin, J.-F. and Théroux, G. (1979). 1976 Census Parametric Evaluation – Reverse Record Check Methodological Report, Part Two. Statistics Canada.

Ha, B., Mayda, M. and Tourigny, J. (1998). Methodology of the 1996 Automated Match Study). Internal Report. Statistics Canada.

Mason, K.E. and Ashraf, A. (1969). Reverse Record Check Sample Design and Estimation Procedure. Dominion Bureau of Statistics.

Mayda, M. (1998). 1996 Census Data Quality Automated Matching Programs, Draft. Internal report. Statistics Canada.

Morel, J. and Farr, H. (2007). Measuring Person Duplication: The 2006 Canadian Census Overcoverage Study. Proceedings of the Section on Survey Research Methods, American Statistical Association, 3380-3387.

Muirhead, R.C. (1969). Reverse Record Check. Sampling and Survey Research Staff, Dominion Bureau of Statistics.

National Research Council. (2008). Coverage Measurement in the 2010 Census. Panel on Coverage Evaluation and Correlation Bias in the 2010 Census, R. Bell & M. Cohen (Eds.). Committee on National Statistics. Washington, DC. The National Academies Press.

Royce, D. (1992). A Comparison of Some Estimators of a Set of Population Totals. Survey Methodology, 18, 109-125.

Statistics Canada. (1980). 1976 Census of Canada Quality of Data – Series 1: Sources of Error – Coverage. Catalog 99-840.

Statistics Canada. (1990). Users Guide to the Quality of 1986 Census Data: Coverage. Catalog 99-135E.

Statistics Canada. (1994). 1991 Census Technical Report – Coverage. Catalog 92-341E.

Statistics Canada. (1999). 1996 Census Technical Report: Coverage. Catlg. 92-370-XIE.

Statistics Canada. (2004). 2001 Census Coverage Technical Report. Catalog 92-394-XIE.

Statistics Canada. (2007). Population and Family Estimation Methods at Statistics Canada. Catalog 91-528-XIE.

Statistics Canada. (2010). 2006 Census Technical Report: Coverage. Catalog 92-567-X.

Théberge, A. and Liu, W. (2002). Non-response Adjustment and Variance Estimation for the 2001 Reverse Record Check. Statistics Canada internal working paper.

Théberge, A. (2006). The 2006 Reverse Record Check Sample Allocation. Survey Methodology, 32, 77-85.

Théberge, A. (2008). Non-response Adjustment for the 2006 Reverse Record Check. Internal report. Statistics Canada.