

MEASUREMENT AND INNOVATION IN THE 2001 CENSUS COVERAGE STUDIES

Colleen Clark, Mark Armstrong and Christian Thibault ¹

ABSTRACT

The Canadian Census of Population is conducted with full recognition that some eligible persons will not be enumerated and that some persons will be enumerated in error or enumerated more than once. As these errors cannot be totally eliminated, an objective of Census collection is to minimize coverage error. The measurement of coverage errors is done independently of Census collection. Several sample surveys and studies are conducted to determine the number of persons who were missed and should have been included in the Census, and the number of those who were counted but should not have been. Estimates of undercoverage and overcoverage are produced for several geographies and demographic groups. Results of the census coverage studies impact population estimates produced by Statistics Canada's Population Estimation Program. Several significant changes in the collection, processing and estimation methodologies were implemented for the 2001 coverage studies. Much of the impact of these changes was in reducing measurement error and increasing efficiency.

KEY WORDS: Census, Census data quality, Coverage error, Net undercoverage, Overcoverage, Population estimation, Reverse Record Check, Undercoverage.

RÉSUMÉ

Le recensement de la population du Canada est effectué tout en sachant que certaines personnes visées ne seront pas dénombrées et que d'autres personnes seront dénombrées par erreur ou plus d'une fois. Un objectif de la collecte du recensement est de minimiser ces deux types d'erreur vu qu'il est impossible de les éliminer complètement. La mesure des erreurs de couverture est faite indépendamment de la collecte du recensement. Plusieurs études et enquêtes à échantillonnage sont conduites pour déterminer le nombre de personnes qui ont été omises et qui auraient dû être incluses dans le recensement et le nombre de personnes qui ont été comptées et qui n'auraient pas dû l'être. Les estimations de sous-dénombrement et de surdénombrement sont produites par régions géographiques et selon les caractéristiques démographiques. Les résultats des études de couverture du recensement ont un impact sur les estimations produites par le programme d'estimation de la population de Statistique Canada. Plusieurs changements importants des méthodes de collecte, de traitement et d'estimation ont été mis en place pour les études de couverture de 2001. L'impact principal de ces changements a été de réduire l'erreur de mesure et d'en augmenter l'efficacité.

MOTS CLÉS : Contre vérification des dossiers; erreur de couverture; estimation de population; qualité des données du recensement; Recensement sous-dénombrement; sous-dénombrement net; surdénombrement.

1. INTRODUCTION

The Canadian Census of Population is conducted with full recognition that some eligible persons will not be enumerated, some will be enumerated in error and others will be enumerated more than once. As these errors cannot be entirely eliminated, an objective of census collection is to minimize these coverage errors. The goal of the census coverage studies is to measure census coverage error. Similar to the Canadian Census, census coverage studies follow a five-year cycle. The development of the coverage studies takes place between censuses with data collection following census collection. Results of the coverage studies are released about two years after Census Day. Compared to the 1996 coverage studies, several changes in the data collection, processing and estimation methodologies were introduced for 2001. Mainly, the 2001 studies employed more technology in survey data collection and data processing. Section 2 presents an overview of Statistics Canada's population estimation methodology for deriving the base population from Census counts and the results of the coverage studies. Section 3 gives definitions of coverage concepts and a summary of

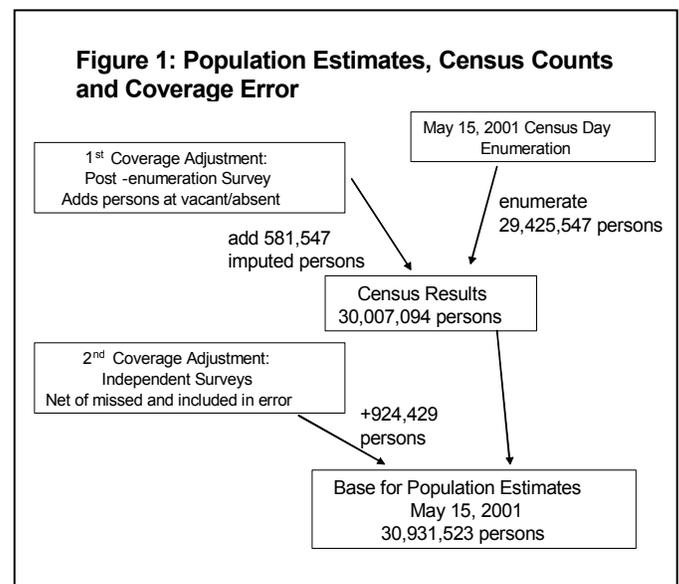
¹ Colleen Clark (colleen.clark@statcan.ca), Mark Armstrong (mark.armstrong@statcan.ca) and Christian Thibault (christian.thibault@statcan.ca), Social Survey Methods Division, Statistics Canada, 15th floor R.H. Coats Building, Ottawa, Ontario Canada K1A 0T6

the different census coverage studies. Section 4 looks at the main study for assessing coverage – the Reverse Record Check. Section 5 presents some methodological issues related to the current coverage studies while conclusions and directions for the 2006 Census coverage studies are given in Section 6.

2. POPULATION ESTIMATES AND COVERAGE STUDIES

The last Canadian Census of Population took place on May 15th, 2001 and about 29.4 million persons were enumerated. Following census collection, several coverage studies are done to assess the extent of missed persons, persons who were counted more than once, and persons who were counted but should not have been. The first coverage study is a post-enumeration survey, the Dwelling Classification Study (DCS). The DCS takes a sample of about 40,000 dwellings classified by Census enumerators as either vacant or occupied but no contact was made with household members. The purpose of the DCS is to estimate the number of persons living in these dwellings on Census Day who were missed by census collection. The result of the 2001 DCS was that about 582,000 persons were added to the census database of enumerated persons. Adding this estimate of missed persons raised the population to just over 30 million persons. Figure 1 summarizes the methodology for deriving the base population for population estimates from Census counts and the results of the coverage studies.

Later in 2001, three surveys independent of the 2001 Census were conducted to measure undercoverage and overcoverage across Canada. Estimates of the number of persons missed (undercoverage) and enumerated in error (overcoverage) were produced and evaluated. For the 2001 Census, the net difference of undercoverage and overcoverage was estimated to be 924,429 missed persons. The results of the coverage studies do not impact the 2001 Census count of about 30 million persons. That is, persons are not ‘added’ to the census database as is done for DCS. Rather, the estimates of net undercoverage are used to produce population estimates via the base population for Statistics Canada’s Population Estimation Program. The adjusted population count is 30,931,523 persons. (Population estimates also make allowance for incompletely enumerated Indian Reserves. Including people missed here, the adjusted 2001 Census count is 30,966,062).



3. COVERAGE CONCEPTS AND COVERAGE STUDIES

There are two components of census population coverage error. ***Undercoverage relates to persons who should have been enumerated on Census Day but were not.*** Persons may be missed for many reasons including an intentional refusal to complete a census questionnaire, being unaware or uncertain that they should count themselves or someone else, or a field collection or processing error occurred. Field collection errors include dwellings completely missed because the enumerator did not know that the dwelling existed, assuming a dwelling was vacant when it was not, or not recognizing sub-units within a dwelling.

Overcoverage refers to persons who were enumerated but should not have been. Persons can be enumerated in error for many reasons including persons who were deceased on Census Day but were enumerated, persons in Canada as non-residents who were enumerated, persons listed on more than one census questionnaire such as those who move around Census Day or those who are uncertain as to where to enumerate themselves.

Separate estimates are produced for undercoverage and overcoverage. In Canada, undercoverage is a more frequent phenomenon than overcoverage. ***Hence, coverage error is often expressed as the net difference of undercoverage less overcoverage, referred to as ‘net coverage error’ or, more commonly, ‘net undercoverage’.*** Coverage errors vary by geography and by demographic group. Net undercoverage, for example, has always been high for young adult males.

3.1 Uses of Coverage Error Estimates

Estimates of coverage error and information on the methodology of the coverage studies serve a number of important purposes including, but not limited to, the following:

- informs users of census data about data quality. As coverage error differs by characteristics such as age and sex, users need to be aware of the extent of undercoverage and overcoverage.
- provides feedback to census collection operations on weaknesses in field collection that can lead to coverage error.
- via population estimates (Section 2):
 - impact on the distribution of federal transfer payments to the provinces.
 - calibrate other surveys such as the Labour Force Survey.

Consequently, the coverage study methodologies are subject to a high degree of scrutiny from internal and external users. Potential bias is a principal concern, especially for federal, provincial, and territorial government representatives concerned with the impact of coverage estimates on transfer payments.

3.2 Coverage Studies

The Automated Match Study (AMS), the Collective Dwelling Study (CDS), and the Reverse Record Check (RRC) measure overcoverage while undercoverage is measured exclusively through the RRC. Section 4 gives an overview of the RRC.

The AMS measures overcoverage by matching the census database to itself. Pairs of similar households containing two or more persons are formed and the members of each household in the pair are compared to one another. A sample of possible matched households is then examined to determine if the same individuals are appearing twice on the census database. For the 2001 AMS, a sample of about 17,000 pairs of households, from a population of about 1,142,000 pairs, was examined.

The CDS is a small study that addresses persons enumerated in both a non-institutional collective dwelling such as work camps or rooming houses, and in a private dwelling. During the census, persons in non-institutional collective dwelling were asked to provide an alternative household address. The CDS verifies a sample of these addresses to determine if the selected person was enumerated at the private dwelling.

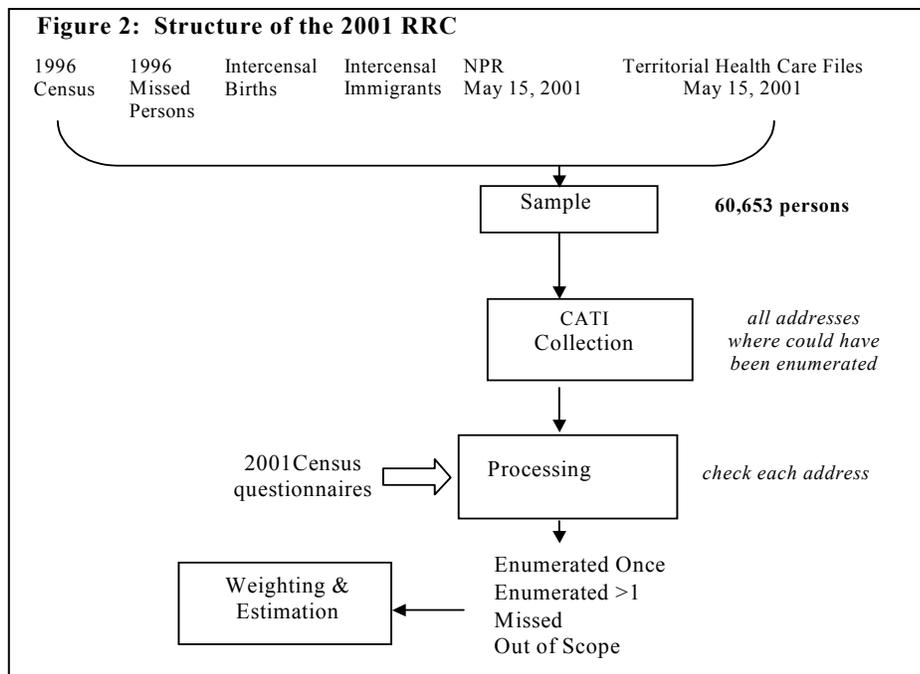
There is no overlap among the three studies leading to double or even triple counting of overcoverage. If overcoverage is detected by the AMS or the CDS and the RRC, this case is removed from the RRC study and reported only by the AMS or the CDS. The total population overcoverage in 2001 is proportioned as follows: 51% AMS, 48% RRC, 1% CDS. This split varies considerably among the provinces and territories.

Sample sizes for all of the studies are sufficiently large to produce very good estimates of net coverage error at the national level, for provinces and territories, and for national age and sex groups. Coverage errors and their standard errors are additionally produced for several geographies and other demographic groups. Given the importance of the estimates as described above, comparing and reconciling RRC estimates with other sources of data is a key part of evaluation. The estimated number of deceased persons, for example, is compared to counts from provincial vital statistics files.

4. REVERSE RECORD CHECK

The RRC was developed for the 1961 Census and the basic premise remains the same. However, with a greater understanding of coverage error and innovations in technology, the 2001 RRC differs from the previous surveys. The RRC is both a sample survey and a study. There is a survey component that has a sample of 60,653 persons. The study component involves work that is done at Head Office in Ottawa and concentrates on those persons that were not contacted during RRC data collection. Figure 2 shows the structure to the 2001 RRC.

The 2001 RRC sample was drawn from six frames independent of the 2001 Census. Territorial health care files provide the frame for the target population of persons who should have been enumerated in the census in the three territories (12% of the RRC sample). For the provinces, 73% of the RRC sample was drawn from the 1996 Census database. The 1996 RRC sample provides a sample of persons missed in the 1996 Census which accounts for 4% of the 2001 RRC sample. Persons born after the 1996 Census and before the 2001 Census are sampled from provincial birth registry files (5% of the RRC sample). Immigrants arriving after the 1996 Census and before the 2001 Census and non-permanent residents (NPR) who possess a permit to reside



in Canada or were claiming refugee status on Census Day are included (4% and 2% respectively of the RRC sample). Frame overlap was addressed throughout the RRC and rules were established for allocating an individual to one frame over another in cases where overlap was detected.

Persons from each frame were selected using a stratified simple random sample design with stratification varying by frame. Variables such as sex, age, year of birth, year of immigration, type of NPR permit and urban/non-urban area are all part of the sample design. The desired level of precision of the estimates was a determinant in the sample size.

In the 2001 RRC, a computer-assisted telephone interview (CATI) was used to gather the survey data. The main purpose of the interview is to determine if the selected person should have been enumerated by the Census and to collect all addresses where they could have been enumerated. The CATI application was BLAISE-based and adopted many of the standards for Statistics Canada telephone surveys. Data collection took place over a ten-month period in order to achieve a very high response rate. About 53 person-years were used to contact and interview persons in the RRC. The overall RRC response rate for the 2001 RRC was 89.8%, about 4% less than in 1996.

Another 30 person-years were used to process the CATI interview data. Processing involved checking 2001 Census questionnaires for each collected address to see if enumeration had taken place there for the sampled person. Sampled persons may have been found at none of their addresses, at one address, or at more than one address. About 300,000 addresses were processed and about 75,000 images of Census questionnaires were examined. The use of images and automated systems for processing staff was an important change from 1996 when paper Census questionnaires and documents were used. The automation enabled processing to focus much less on document management. Following processing, each sampled person was classified as enumerated once, enumerated more than once, missed (should have been enumerated but was not), a non-respondent, emigrated, deceased, or otherwise out of scope.

The final stage of the RRC is the weighting and estimation. This involved a non-response adjustment procedure to redistribute the survey non-respondents to a group of similar respondents. The 2001 variance calculations and non-

response adjustment methodologies differed from the 1996 RRC in several respects. Principally, the variance estimation was improved as a result of moving away from a replicate procedure to the Statistics Canada Generalized Estimation System (GES). The 2001 non-response adjustment was based on very precise definitions and a sequential five-step adjustment procedure.

4.1 Automation in the 2001 RRC

Automation in data collection and processing introduced for 2001 resulted in a reduction in survey measurement error and a more efficient survey. Current desktop tools facilitated the production and dissemination of coverage error estimates while the use of Statistics Canada's GES enabled derivation improvements to the estimated standard errors.

4.1.1 Data Collection

Data collection moved from the paper and pencil interview (PAPI) mode used for the 1996 RRC to computer-assisted telephone interviewing (CATI). Quality improvements realized with CATI included expanded content. Blocks of questions applicable only to some respondents were added by implementing automated skip patterns that brought the interviewer to the questions only when the respondent met the programmed criteria. As well, CATI allowed edits to be programmed that identified the consistency of responses to interviewers. CATI also customized each interview by pre-filling fields where applicable such as the sampled person's name. The CATI environment allowed interviewers to concentrate more on communication with the respondent rather than shuffling papers or turning pages.

Due to outdated or missing addresses and telephone numbers, locating selected persons was a significant part of the RRC. A CATI tracing module to record tracing activities gave interviewers more comprehensive information on locating attempts and tracing sources than was the case for the 1996 RRC.

Field management of the RRC also improved as a result of CATI management tools available to interviewers and senior field staff. Management information reports were critical in identifying the need for focused collection strategies. CATI also improved RRC data processing by allowing processing to immediately follow collection since questionnaires did not need to be captured. Also, capture error was greatly reduced as no separate capture of paper questionnaires was required.

CATI data quality improvements were realized only with a substantial investment in developing a CATI environment that would work well for interviewers in all situations. A valuable lesson was that the time required for writing specifications and then programming and testing the CATI application requires significantly more resources and a much earlier start date than had been the case for a PAPI RRC.

4.1.2 Data Processing

Past processing of RRC data required a lot of paper-based operations including the retrieval of Census questionnaires and other documents. Automated operations and electronic documents introduced for the 2001 RRC meant staff focused on research rather than on managing paper. Images of Census questionnaires, for example, were quickly consulted rather than the longer task of locating paper questionnaires. Automation also permitted a better flow of cases between the staff responsible for different operations. Eliminating paper and their requisite filing systems meant that cases were in transit between operations for less time. Case management reports were also easier to produce by the processing supervisors themselves.

There were more addresses collected in the 2001 RRC. Processing of the increased volume was achieved because addresses could be processed faster than ever before.

Automation is believed to reduce item non-sampling error in the steps of processing in the same way as CATI enabled edit failures to be managed at the point of interviewing. One measure of the quality improvements due to automating processing was that only about 1% of all the addresses collected could not be resolved, a decrease from 2.8% in the previous RRC.

4.1.3 Other Improvements

Production of the estimates of coverage error was a lengthy process in the 1996 RRC. In 2001, the use of current desktop tools meant a much shorter and more efficient process thereby allowing more time for evaluation, analysis, and developing new tables. SAS and MS Excel were the principal tools for producing point estimates. Another improvement was in variance estimation where a direct formula was used via Statistics Canada's Generalized Estimation System. In the past, custom SAS code carried out variance estimation by a replicate approach with five replicates. Dissemination was also more efficient with main users accessing a Statistics Canada server to download tables and associated documentation. A notable benefit of the efficiencies gained in production and dissemination was that estimates for different estimation scenarios could be produced in a short period of time.

5. METHODOLOGICAL ISSUES

Although sampling errors tend to be small for the high levels of geography and demographic groups, there is concern about potential bias. Because coverage study results impact federal transfer payments, bias should be minimized. One important example of potential bias is the RRC adjustments for non-respondents - selected persons not contacted and those for whom an interview was completed but the addresses were too vague for staff to locate a Census questionnaire. As the non-response rate increases, as it did for the 2001 RRC, there is concern that the impact of any bias may be increasing.

Overcoverage in the 2001 Census was less than 1% of the population, about 298,000 persons. Due to the rarity of overcoverage and the challenges in finding efficient detection methods, there is some potential for bias.

6. CONCLUSIONS AND LOOKING TO 2006

The Canadian Census of Population has coverage error resulting from not counting persons who are eligible for inclusion and from counting others who should not be. Coverage studies measure overcoverage and undercoverage using small and large sample surveys. Study results are combined to produce estimates of coverage error for various geographies and demographic groups. These estimates are incorporated into Statistics Canada's Population Estimates Program. Although many innovations reducing survey measurement error were introduced for the 2001 coverage studies and non-response increased in the RRC, the increase in net undercoverage recorded by the coverage studies is attributable to census collection and not to changes in the coverage studies.

The 2006 Census will be conducted quite differently than past censuses. Changes include the use of a mail-out procedure from a national address register in the urban centers, internet responses, and a centralized follow-up operation. The accurate measurement and understanding of coverage error is therefore even more important. Incorporating efficiencies in survey design and managing the cost of developing and conducting the 2006 coverage studies remain important considerations.

ACKNOWLEDGEMENTS

Census coverage error can only be determined by independent and detailed studies. Although the team involves many people in Statistics Canada, special thanks are given to Clément Brunet, Norm Crampton, Peter Dick, David Dolson, Heather Farr, Harold Goodwin, Gildas Kleim, William Chunxiao Liu, Josée Morel, Michel Parenteau, Heather Richards, Alain Théberge and Carol Wolkowski. The work of Rob Lethbridge and the national RRC CATI collection team, Luc Tremblay and the CATI development team, along with Ann Charron and the efforts of the Head Office processing team are also respectfully acknowledged.

REFERENCES

Statistics Canada (1999). *Coverage: 1996 Census Technical Report*. Catalogue No. 92-370-XIE.