Combined Census Research in Canada: 2021 Census of Population Statistical Contingency Plan

Georgina House, Jean-François Simard, Karelyn Davis

Statistical Integration Methods Division, Statistics Canada
100 Tunney's Pasture Driveway, Ottawa, ON Canada
Corresponding author email: georgina.house@statcan.gc.ca

Abstract

Statistics Canada has been conducting research on the use of administrative data to support, supplement, or replace traditional census collection for the last 10 years. In the context of this research, a combined census methodology which involves combining traditional census collection methods with administrative data aims at uniting the best information from both sources. Due to collection issues related to the COVID-19 pandemic, a statistical contingency plan was implemented in the 2021 Census based on these principles and using existing combined census research. Simulations of prospective combined census strategies are being conducted using an administrative sociodemographic database and the 2021 Census response data. Several criteria are being considered when evaluating the simulations, including data quality, geography, dwelling type, coverage, and traditional census collection methods. This presentation will discuss the methodology behind the simulations as well as some analyses of the results, specifically a comparison of benchmarks including traditional census counts and demographic estimates. It will also discuss how the statistical contingency plan was implemented at the imputation stage after the 2021 Census collection ended, as well as future research.

Key words: Administrative Data, Imputation, Canadian Census, Simulations

1. Introduction

Due to public health restrictions from the COVID-19 pandemic which had the potential to reduce or prohibit collection for the 2021 Census, a statistical contingency plan was conceived in March 2020 and later implemented in the 2021 Canadian Census. Prior to the pandemic, Statistics Canada had been conducting research on the use of administrative data to support census collection. This research was based on administrative data sources (Statistics Canada, 2019), which was leveraged in the planning for a contingency. Simulations of prospective contingency strategies were conducted using an administrative sociodemographic database.

Canadian Census

The Canadian Census of Population is a de jure census and occurs every 5 years. The census was last undertaken on May 11, 2021, and the next census will be in May of 2026. The traditional census is based on collection from all dwellings in Canada and includes a mandatory response. Canadian households receive one of two versions, the short or long form questionnaire. The short form contains basic demographic variables such as age, sex, language, number of usual residents, and household composition and is distributed to 75% of households. Nationally the long form is provided to a 25% sample of households and contains more information, including questions on the number of rooms in the residence, education, ethnicity and income information (Statistics Canada, 2022). The results are used in many ways, including determining electoral representation to Parliament, calculating transfer payments between levels of government, and supporting various government programs. Individuals are

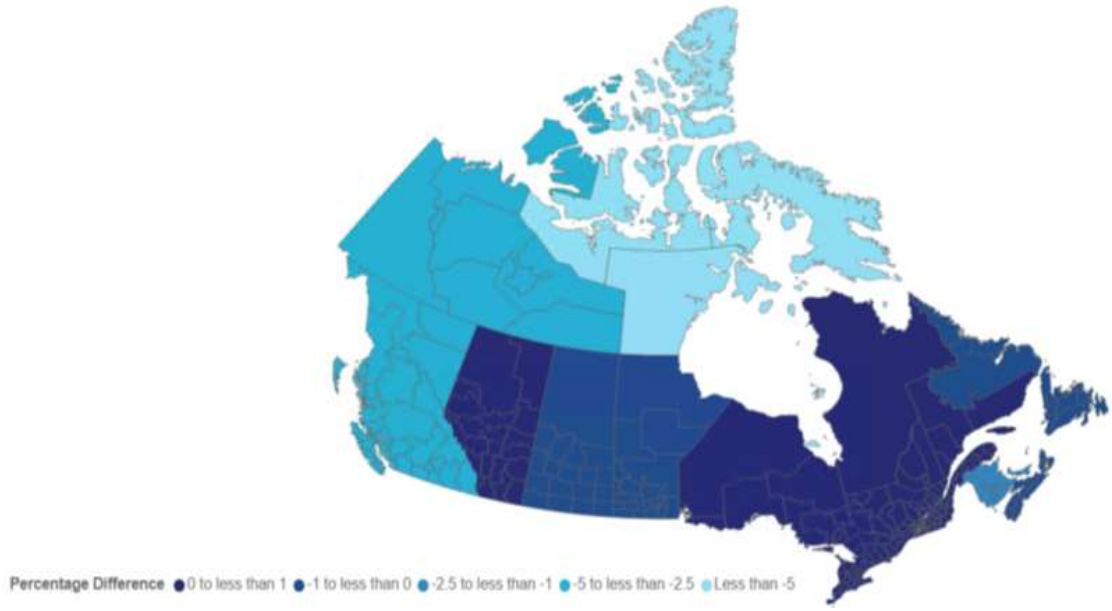
enumerated at their usual place of residence in Canada. The entire population of Canada is enumerated including Canadian citizens, landed immigrants and non-permanent residents, as well as Canadian citizens and landed immigrants who were temporarily out of the country on Census Day.

Combined Census Research

Combined census research has been ongoing at Statistics Canada for several years (Royce, D, 2017). The census program in 2011 consisted of the Census of Population and the National Household Survey which was made voluntary for this year due to a governmental decision. The voluntary nature resulted in a lower-than-normal response rate which helped to demonstrate the need to research alternative ways to look at census collection. Research into a combined census – a census combining traditional survey collection and administrative data sources - has been conducted in Canada due to multiple reasons. For instance, the traditional census is subject to unforeseen events like natural disasters, labour disruptions and elections, that can seriously affect census fieldwork and, in some cases, limit the possibility of recovery. In 2016 there were significant wildfires in Fort McMurray Alberta, Canada that made census enumeration impossible (Statistics Canada, 2019, Appendix 1.4). In the Fort McMurray area previously completed online questionnaires were supplemented with information from linked administrative files (for name, date of birth, sex and marital status). Natural disasters have occurred recently, with both floods and fires causing collection difficulties in the past two census cycles. This research also allows us to look to the future and a move to a combined census making use of administrative data, as has occurred in other countries (UNECE, 2018).

Prior to the pandemic, the 2016 Statistical Demographic Database (SDD) was the second phase of a pilot project that gathered administrative data in a comprehensive file. This research was instrumental during the Fort McMurray challenges in Census 2016. As part of the combined census research, empirical simulations were completed which focused mainly on splitting up the data into two groups: private mail-out dwellings and all other types of dwellings. In a traditional census, private mail-out dwellings receive, by mail, an invitation letter to complete the questionnaire online. This letter contains a secure access code and a telephone number to allow the respondent to request a paper questionnaire if they prefer. Private mail-out dwellings are the dwellings with the best administrative data because the address quality of mail-out dwellings tends to be more accurate. Private mail-out dwellings were then replaced by administrative records under different circumstances. In some scenarios all private mail-out dwellings were enumerated using administrative based information. In other scenarios the use of administrative data was determined based on the coherence rate within each municipality. The coherence rate was calculated based on the consistency of household composition between the administrative data compared to the census data. The use of administrative data was also based or where the address-missing rate was above a certain threshold. The address missing rate was calculated according to the capacity of assigning geography to administrative data within a geographic area. Figure 1.1. uses a map of Canada to show some results from one of the empirical simulations. The colours represent categories of relative differences, as calculated when comparing the simulated provincial and territorial estimates to the demographic estimates, adjusted for census net coverage errors. The blue scale goes from the darkest being closest to zero to the lightest being furthest away from zero. A relative difference of zero would be ideal. We can see that most of the provinces have relative differences close to zero, indicating the combined census is promising for producing provincial estimates. There is ongoing research to evaluate the provinces and territories with larger relative differences.

Figure 1.1. Canada: Percent Difference Simulation Estimates Compared to Demography Estimates: 2016 Results



2. Description of the Census 2021 Contingency Plan

The census contingency plan was conceived following the onset of the COVID pandemic in 2020. The team wanted to be prepared for any potential issues with the 2021 Census. The combined census research was leveraged as an input for testing scenarios for the contingency plan. Scenarios under investigation and methods used during the research were considered as an input for a test to use administrative data for direct replacement in 2021 to mitigate any potential nonresponse issues.

The initial plan was for administrative data to be used for direct replacement at the collection stage of the census where administrative data was of sufficient quality. During the contingency plan, exploration and adaptation were necessary. Given the unknown impact of the pandemic on response rates, use of administrative data was reserved for post-collection imputation. For reasons of both production and social acceptability, it was decided to use administrative data in the imputation process to improve data quality in areas of low response rates.

For the 2021 Census an administrative database was derived, using probabilistic linkage of tax data, vital statistics, immigration data, driver's licenses, social insurance information as well as other information. This file, denoted the Statistical Demographic Database (SDD), is based on administrative government files used in other programs at Statistics Canada. The SDD contained 52 million administrative individuals, which is larger than the Canadian population of approximately 38 million people. Statistical models were used to identify individuals within the census target population. Models were also used to determine the usual place of residence (called the household model). In total 37 million individuals were found to be in scope. There was good accuracy at the national and provincial levels with 86% national accuracy, where 86% of SDD individuals were assigned to the same dwelling as the census. In this paper, we focus on the simulations and implementation of the statistical contingency plan. More information on the derivation of the statistical models can be found in Lundy (2022) and Yoon et. al. (2022).

2.1. Contingency Scenarios

During the lead up to the implementation of the contingency plan, several scenarios for administrative imputation were tested using Census 2016 data, and the results examined. Developing the scenarios to test the contingency was a challenge. The requirements during the scenarios had to take into account uncertainties, such as the final response rate. Due to these uncertainties different scenarios were considered.

The first set of contingency scenarios involved 'simulating' nonresponse for a certain percentage of dwellings which responded to the 2016 Census. These scenarios were performed to simulate a scenario where collection might have had to stop early. They included dwellings which responded to the census, and other resolved cases (with unoccupied and cancelled dwellings) along with non-respondent dwellings. While the official census date was May 11th, the simulation chose other dates to represent the last respondents. Three of these scenarios were considered: i) the last 2% of dwellings on July 15, ii) the last 5% on June 30th and iii) the last 10% of respondents occurring on or after June 17th in the 2016 Census. We simulated the additional nonresponse by blanking out responses coming from these three groups. It is worth noting that the 2016 Canadian Census had a response rate at the dwelling level of over 98%.

The next scenario considered the impossibility of doing in-person nonresponse follow-up (NRFU). Dwellings with no phone number at the beginning of NRFU that were resolved by field staff were set to be nonrespondents for this scenario. This created a nonresponse rate of around 8%.

Finally, the last scenario considered additional nonresponse in urban areas, given higher COVID-19 infection rates in cities. Increased nonresponse was simulated in large census metropolitan areas (Montreal, Toronto, Calgary, and Vancouver) as they were seen as being more affected by the pandemic. Dwellings resolved on or after June 1st (start of nonresponse follow-up) were set as non-respondents. The scenario was conducted to illustrate a case where no nonresponse follow up was possible in the urban areas.

2.2. Administrative Data

During the tests, administrative data was used in several different ways. We considered using all available administrative data, along with scenarios where there was a cap on the number of administrative households that were used, for example only three households per collection unit. The collection unit is the primary collection geographic area used to assign and monitor work in the various collection activities. (Statistics Canada, 2022) We also considered the level of quality of the administrative data, testing different scenarios with different levels of quality. The building of administrative households and their level of quality was completed by statistical modelling.

Administrative households were derived using researched methodology (Lundy, 2022, Yoon et al, 2022). This research noted administrative records varied in quality with some over-coverage of individuals. A modified Euclidean distance function was used to rank records in terms of the individual and household-level coherence with Census 2016. In the end, administrative households with the best quality were used in the contingency scenarios. Households defined as 'high' and 'medium' quality were considered in-

scope for the simulations. Out of 14.2 million admin households, 9.6 million were deemed eligible for direct imputation (67.6%).

2.3. Final Scenario: Whole Household Imputation

As mentioned previously, due to the production system limitations as well as some social acceptability concerns, it was seen as more prudent and acceptable to use administrative data during post-collection imputation. The previous contingency research was considered and combined to proceed with a new scenario. This final scenario was performed to incorporate administrative data into Whole Household Imputation (imputation for total nonresponse) of the census short-form characteristics. When simulating this scenario, we considered nonresponse for the last 10% of responding households as well as some increased nonresponse in certain areas due to potential pandemic nonresponse. This targeted nonresponse was in certain areas and included two urban areas and two rural areas. The areas of interest were the cities of Edmonton and Toronto, along with rural Newfoundland and rural Saskatchewan. The increased nonresponse in these areas was based on dwellings with no phone number that had been resolved by field staff during nonresponse follow-up.

There were three contingency scenarios run through the whole household imputation. CANCEIS, which is a Statistics Canada generalised system, is used for donor imputation in the Canadian Census. Age, sex and household size were imputed with administrative dwellings of high and medium quality. One of the main objectives was to assess whether the consistency between the 2016 and 2021 Census estimates was maintained. The initial contingency scenario (CS0) was a run of traditional donor imputation, using the same non-respondent dwellings from the 2016 Census but with 2021 geographic information. We use this as a base for comparison. The second contingency scenario (CS1) included the extra nonresponse from the final scenario discussed previously using the traditional whole household imputation method for all non-respondents. And finally, the last contingency scenario (CS2) included the extra nonresponse and used the links to the administrative data when available. Administrative data was used to impute age, sex, and number of residents.

3. Results

Figure 3.1 shows the imputation rates comparing contingency scenarios CS0 and CS2 by province. The original imputation rate is close to 2% in most of the provinces, with higher rates noted in the territories. The rate ranges from 7 to 16% in the contingency scenario CS2.

Figure 3.1: Imputation Rates: Contingency Simulations

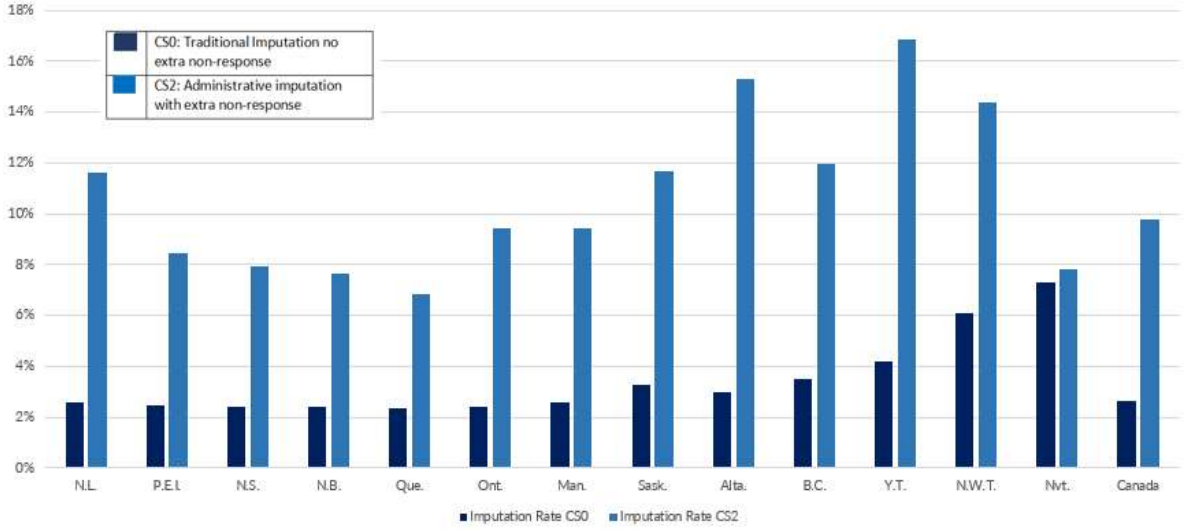
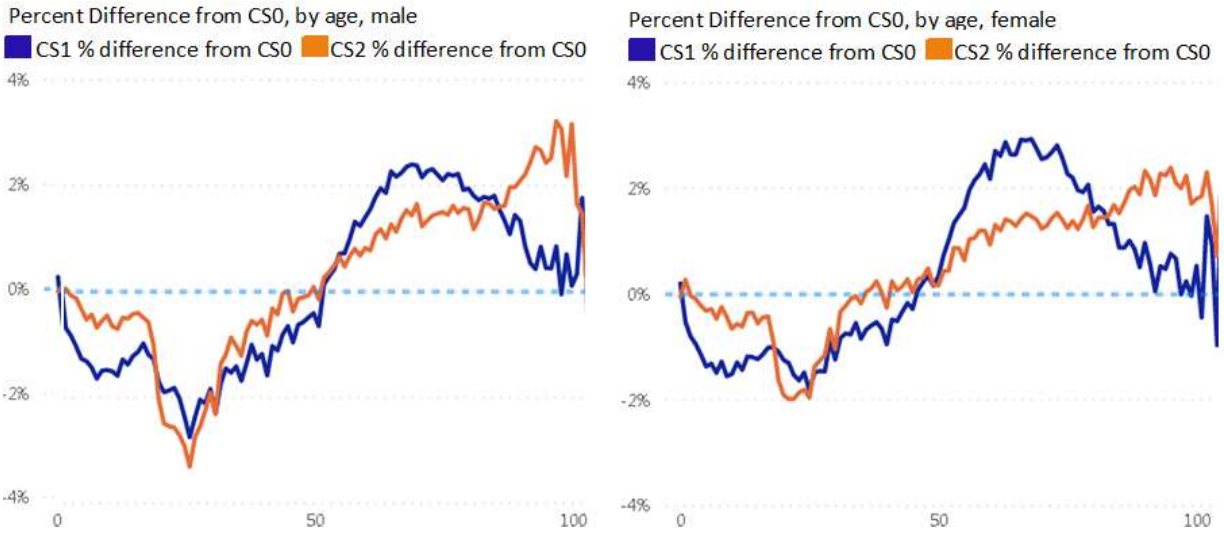


Figure 3.2 represents the percent difference of population estimates by age and sex between contingency scenario 0 (or the baseline) and scenario 1 (using WHI on all nonresponse) or scenario 2 using the administrative data. The orange line represents admin donor imputation, whereas the blue line represents traditional donor imputation. Values closer to the zero axis are best, as they are the smallest difference from the baseline scenario. Here we can see that the administrative imputation does better for children and individuals aged 30 – 80, but worse for individuals 18-29 and 80+. Additional research is planned for these age groups.

Figure 3.2: Percent Difference by Age and Sex



4. Contingency Plan Implementation

Data quality, computational limitations, operational and system limitations, as well necessity and proportionality were considered in order to decide how to implement the contingency plan. The decision was based on sound methodology, as demonstrated through the simulations presented in the previous

section. The main portion of the contingency scenarios studied were performed on the last 10% of data. The analyses that were performed showed that the use of the contingency had lower mean absolute percentage error as compared to the demography estimates with scenario CS0 when the response rates were around 90%. The administrative contingency plan was therefore implemented in collection units with a response rate of less than 90%. A response rate of 90% is considered a relatively low response rate in the Canadian Census. Despite the challenges of the pandemic the 2021 Census had a 98% response rate nationally. Nevertheless, there were 9.2 million dwellings eligible for administrative replacement in Census 2021 using the research described in this paper.

From Table 1, we note the counts of the records used in the final contingency for Census 2021. Administrative data was used for around 10,000 households or 22,000 people. The amount of data used was quite low compared to the whole population, but comparing it to the whole household imputation counts, it is a bit more substantial. In hindsight more could have been used, for example we could have considered all collection units instead of just those with less than 90% response rate. This approach will be considered for future research.

Table 1: Census 2021 Contingency Counts

	Rounded Counts
Non-respondent households	122,000
Administrative households	10,600
Number of administrative people	22,500

5. Future Considerations and Limitations:

Due to the nature of the contingency, some limitations were present. Limitations included employing a new process, using administrative data, challenges in integrating geography, and timing. The edit and imputation system had a limited number of runs available due to the timing of the studies relative to the production window and to computing availability. This limited the possibility of testing several options. Conversions had to be made from the 2016 geography information to the 2021 geography, which was straightforward for some records, but more complicated for others.

The scenarios simulated a certain kind of nonresponse using the information available. For example, the last 2%, 5% and 10% of the dwellings remaining in the workload were simulated as nonresponse, as well as dwellings with no phone number at the beginning of NRFU that were resolved by field staff. The requirements during the scenarios had to take into account uncertainties, such as the final response rate. Because of these uncertainties, they were hard to define. Future scenarios could include more in-depth census response analyses and respondents/non-respondents' socio-demographics could be considered.

And finally, to inform research in a future combined census, consultations should be pursued between Methodology, Census Futures, and other partners in order to align the combined census testing

strategies with 2026 administrative data support planning, including the upcoming 2024 Census behavioural tests. This approach will help to inform the future of a combined census in Canada.

Combined census research has been ongoing for several years. The contingency plan in 2021 was a means to advance this research for more administrative use. Looking forward, Statistics Canada is planning an increase in the use of administrative data for the 2026 Census, which will be a traditional census with additional administrative usage in nonresponse follow up. Further, research is underway for a potential combined census in 2031, based on international experiences and the Canadian context. At present, three options are under study, ranging from dwelling-based collection to an administrative-first approach, similar to other countries. While Canada does not have a population register, the 2021 contingency analysis noted potential for increased use of administrative data in future Canadian Censuses and greatly advanced the use of such research in census collection and production.

References

Dolson, D. and Dasylva, A. (2017). Census Coverage Assurance and Measurement in an Administrative Data Assisted Census. Report to the Advisory Committee on Statistical Methods: Internal Statistics Canada document.

Laperrière, C. (2020). "Collection Methodology for the 2021 Census and Administrative Data Contingency Plan", Report to the Advisory Committee on Statistical Methods: Internal Statistics Canada document.

Lundy, E. (2022). Predicting the quality and evaluating the use of administrative data for the 2021 Canadian Census of Population. *Statistical Journal of the IAOS*. 38: 1177-1183.

Privacy Impact assessment: Generic Privacy Impact Assessment for Statistics Canada's Statistical Programs (statcan.gc.ca)

Royce, D (2017). First report on the Census Program Transformation Project: Researching a new approach to Census taking URL: https://www12.statcan.gc.ca/census-recensement/fc-rf/98-506-x/98-506-x2017001-eng.cfm

Statistics Canada (2019) Appendix 1.4 – Note describing the Wood Buffalo census subdivision data collection methodology and the use of administrative data sources (statcan.gc.ca)

Statistics Canada. (2021), Guide to the Census of Population (statcan.gc.ca)

Statistics Canada. (2022), Guide to the Census of Population, 2021, Appendix 1.7 – Use of administrative data to impute non-responding households in areas with low response rates (statcan.gc.ca)

Trépanier, J. (2017). Overview of the Census Program Transformation Project, Report to the Advisory Committee on Statistical Methods: Internal Statistics Canada document.

UNEC (2018), Guidelines on the use of registers and administrative data for population and housing censuses | UNECE

Yoon et. al. (2022) Modernization of the Canadian Census: An Administrative Data-Driven Approach to Defining Households. Proceedings of the Survey Methods section of the Statistical Society of Canada, May 2022. URL: https://ssc.ca/sites/default/files/imce/yoon_ssc2022.pdf