Economic and Social Council

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-first Meeting
Geneva, 18–20 September 2019
Item 4 of the provisional agenda
Research on the use of administrative data for censuses

Researching administrative data sources and geospatial datasets for Canada's 2021 census

Note by Statistics Canada*

Summary

> In preparation for Canada's 2021 Census of Population Programme, broad consultations were conducted on census content. The feedback from these consultations identified a number of data gaps that required new content, or revisions to existing content. As a strategy to reduce burden, subject matter teams were asked to investigate administrative sources that could replace questions, add new content or improve existing content without the associated increase in response burden and cost. Having successfully replaced income questions with administrative data in 2016, there was a strong precedent for this approach. Several content areas conducted research to assess replacing existing content and adding new content by integrating other administrative data sources, including immigration and housing data. Complementary to this work, effort has been underway to use administrative data sources to further improve concepts such as commuting to work, and to increase the information on individuals living in collectives, including military bases, nursing homes and prisons. This paper presents the research done in Canada on using administrative data to replace content, add new content and improve existing content for the 2021 Census of Population.

---

* Prepared by Heather Dryburgh.

Please recycle

I. Introduction

1.   In preparation for Canada's 2021 Census of Population Programme, broad consultations were conducted on census content. The feedback from these consultations identified a number of data gaps that required new content, or revisions to existing content.

2.   As a strategy to reduce burden, subject matter teams were asked to investigate administrative sources that could replace questions, add new content or improve existing content without the associated increase in response burden and cost. Having successfully replaced income questions with administrative data in 2016, there was a strong precedent for this approach.

3.   This paper presents the research done in Canada on using administrative data to replace content, add new content and improve existing content for the 2021 Census.

II. Administrative Data

4.   Within census content areas, research is underway to assess the quality, strengths and weaknesses of replacing some questionnaire content with data from administrative sources. These content experts are exploring the potential of acquiring information on immigrant status, year of immigration and housing information from administrative sources.

5.   Complementary to the work of adding new content, efforts have been underway to assess the potential for using administrative data sources to further improve concepts such as commuting to work, and to increase the information on individuals living in collectives, including military bases, nursing homes and correctional institutions.

A. Immigration

1. Replacing immigration questions on the 2021 census

6.   Canada has significant administrative data holdings on immigration, which provide an opportunity to consider how these may be used to replace questions on the census. In the 2016 census, the *admission category* and *applicant type* variables provided information on the conditions under which immigrants have been granted the right to live in Canada. These variables were drawn from Immigration, Refugees and Citizenship Canada's (IRCC) administrative immigrant records. The administrative data for these variables were available only for immigrants who had arrived since 1980. As a result, only persons who reported having immigrated since 1980 in response to the year of immigration question (Question 15) on the census questionnaire were considered "in scope" for replacement with administrative data. Missing values for admission category were imputed for immigrants where an administrative value was not available.

7.   For 2021, Canada is testing replacing the year of immigration and immigrant status using similar methodology and data sources. IRCC administrative data currently available to Statistics Canada include:

  (a)   Detailed immigration records from 1980 to present;

  (b)   Limited immigration records from 1952 to 1979;

  (c)   Non-permanent resident permit records from 1980 to present.

8.   These files provide the actual administrative immigrant status and year of immigration, where applicable, for the immigrants and non-permanent residents they cover. While IRCC data reflect an administrative census of all incoming immigrants and non-permanent residents to Canada, they are not updated to capture deaths or outmigration. Therefore, they cannot be used in isolation to estimate the current population of immigrants living in Canada.

9.   Using similar methods to those which been employed for past censuses (Biernot 2017; Brennan 2013; Brennan 2011), a test of the data integration between census respondents and IRCC administrative data is planned to take place immediately after 2019 census test

collection, and before any edit and imputation processes. This integration of data from two sources will use all common fields available, such as name, date of birth, address, etc. For the test, the response to the immigrant status and year of immigration questions will be replaced by the administrative value and compared. Missing values will be imputed for immigrants where an administrative value is not available.

10.   The citizenship question is fundamentally linked to both concepts (year of immigration and immigrant status), so responses to the citizenship question will be used to determine which individuals require imputation. The Canadian population is divided into four subgroups as shown in Table I below:

Table 1
Immigrant status and citizenship, Canada

Citizenship status	Immigrant	Non-immigrant
Canadian by birth		Canadian by birth
Canadian by naturalization	Canadian by naturalization	
Not a Canadian citizen	Permanent resident	Non-permanent resident

11.   Anyone who is a Canadian citizen by birth cannot be an immigrant. Alternatively, anyone who is a Canadian citizen by naturalization must be an immigrant. Those who do not hold Canadian citizenship are either permanent or non-permanent residents. In certain cases, respondents may say that they are Canadian citizens by birth but be identified in the administrative data as immigrants – such cases will be resolved during the data processing.

12.   The research will also evaluate how replacing the immigration questions could affect historical comparability. An evaluation of the quality of the existing census questions was conducted in order to compare the data quality differences between asking questions and using administrative values. Advantages and disadvantages associated with both options will be fully documented in a report in September 2019 along with 2021 census content recommendations for government.

13.   Additional administrative information to enhance the analysis of immigration that may be added to the 2021 census from IRCC's immigration records includes: *year of arrival; temporary resident type* (work permit holders, study permit holders, refugee claimants, other permit holders); *intended destination; country of last permanent residence*. Consultations with stakeholders are in progress to determine the final list of variables as well as an assessment of the quality of the variables and the operational considerations.

2.   Limitations of using administrative data for immigration

14.   For the purposes of replacing the two immigration questions on the census of population, the coverage of IRCC data introduces three limitations.

15.   First, there are no immigration records available prior to 1952. While immigrants who arrived prior to 1952 are a decreasing population and of lesser interest for researchers, this group still represented 111,000 immigrants (or 1.5 per cent of all immigrants) living in Canada according to the 2016 census. Among them, 86,285 (or 1.2 per cent of the immigrant population) arrived from 1947 to 1951 and the rest (representing 0.3 per cent of the immigrant population) landed prior to 1947.

16.   Second, non-permanent resident records contain information on permit holders only; any accompanying family members who do not hold permits in their own right are not covered. Although some visitors respond to the census questionnaire, visitors are not covered by IRCC administrative data integrated with the 2016 census.

17.   Third, beyond coverage, another limitation arises from the quality of the immigration records prior to 1980. Overall, there is less information included on these records that can be used for data integration purposes. In particular, records from 1961 to 1972 contain an

incomplete date of birth which leads to lower integration rates. Figure I shows the impact of incomplete date of birth data on the degree of success of the integration. There is a notable drop in the integration rate between 1961 and 1972.

Figure I
Percentage of 2016 Census immigrants integrated with administrative immigration records by 2016 Census year of immigration
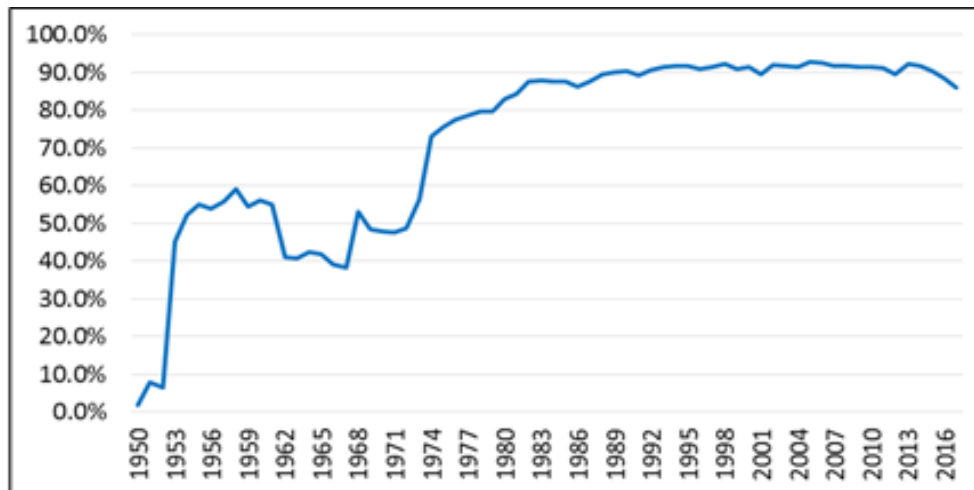
Source: Statistics Canada, 2016 Census integrated with IRCC administrative immigration data since 1952

18.   Additionally, moving away from asking the questions and towards using the administrative values directly means that one data source cannot be used to certify the other, as was the case in the 2016 census. Moreover, replacing the questions with administrative data may affect historical comparability. In particular, shifts in the distribution of year of immigration may occur when self-reported "perceived" year of immigration (e.g. year of arrival) is replaced with actual year of immigration from administrative data. For example, some immigrants may have reported their year of arrival, as they may have resided in Canada as non-permanent residents prior to being admitted as permanent residents.

19.   Coverage of non-permanent residents is less comprehensive than that of permanent immigrants. Non-permanent residents' information is restricted to asylum claimants and temporary residents with a valid permit in 2016.

3.   Research: how to address missing integrations or inconsistencies between integrated values and other census responses

20.   In general, similar edit and imputation methods to those used in the past can be applied. However, in the context of using integrated data, the absence of a value could be the result of a missing integration (or the equivalent of item non-response), or could be reflective of respondents who are out-of-scope for the linkage. Integration of other data sources to mitigate the coverage issues associated with the administrative data is being tested.

21.   In general, the 2016 census and the preliminary results of the administrative data replacement are similar, as seen in Figure II. However, the administrative data replacement results do show some confusion between immigrants' responses and the administrative data. The delayed response effect in the early 1990s caused by a large number of asylum claimants is a good example: respondents provided year of arrival instead of year of immigration.

Figure II
Preliminary test results by single year of immigration, 2016 Census and test results
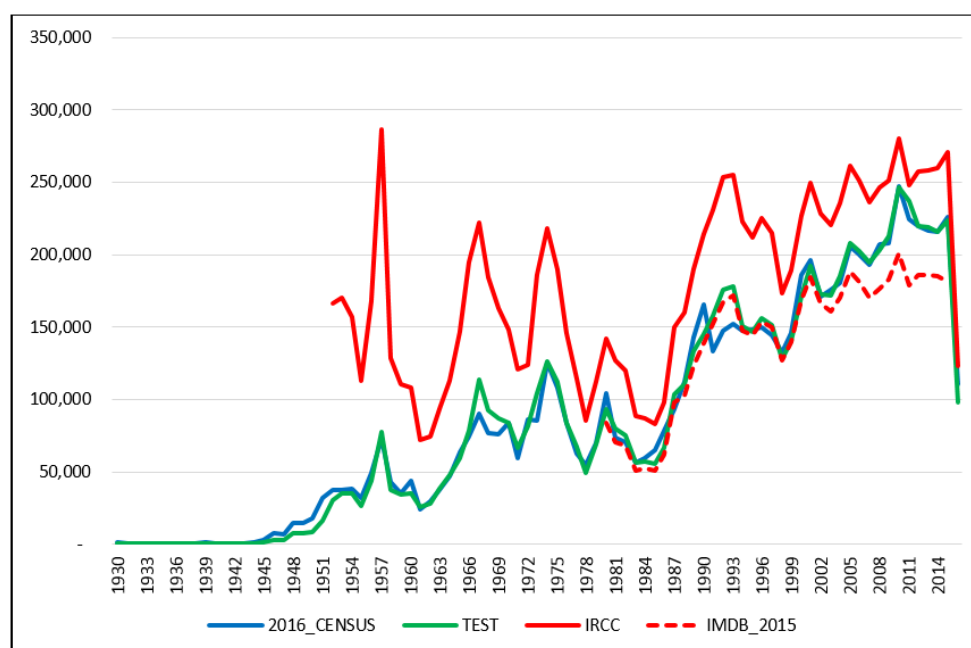
Source: Statistics Canada, 2016 census and administrative data replacement results, Census of Population and Longitudinal Immigration Database (IMDB) for tax year 2014; and Immigration, Refugees and Citizenship Canada (IRCC).

B. Housing

1. Using Big Data for utilities information

22.	In previous censuses, Canadians have been asked a number of questions about their housing, including the amount of money spent on utilities such as electricity. Statistics Canada is researching the feasibility of using electrical supplier information to replace this question on the questionnaire. There are quite a few electricity suppliers in Canada. Statistics Canada has been establishing agreements to acquire billing cycle data from many of them. To date, several utilities are able to supply monthly consumption and billing data to Statistics Canada.

23.	Statistics Canada will evaluate these data for two immediate purposes for the 2021 census. First, the data will be assessed as to whether they could assist Statistics Canada in establishing the occupancy status of a non-responding dwelling. The desired outcome would be to reduce follow-up efforts. Secondly, the extent to which these data could be used as a replacement for the electricity payments question on the census will be evaluated.

24.	The research into integrating or replacing data from utilities providers is in its first phase. As part of the first phase it will be necessary to evaluate how the census questionnaire can be modified to accommodate the fact that full coverage of the country is not currently possible (only a subset of suppliers will be on board by 2021). It has been shown in the 2019 census test that it is possible to adapt the electronic questionnaire to specific areas and random subsets to present one version or another of the questionnaire (with or without specific questions). Preliminary results also show reductions in completion time for households where the electricity question was not asked. However, these gains were partially offset by the confusion introduced for some respondents who may have reported electricity payments in fuel payments or other utility payments. Quality control on these factors will be important.

25.	The second phase will be to bring together electric meters information with census dwellings and assess the discrepancies. In most cases, the relationship is one-to-one and a mapping can be established. However, in other cases multiple meters serve a single dwelling or, alternatively, one meter serves multiple dwellings. Yet others do not provide sufficient

information to identify properly the related census dwelling. These more complicated patterns are concentrated in specific areas or types of dwellings. This phase will assess how to appropriately distribute the consumption among the dwellings.

26. The final phase will be to assess the impact on the results using a mixed-mode approach and anticipate the final impact of such an integration.

2. Limitations to using administrative data for housing

27. The general caveats apply for integrating administrative data in the census. They all must be solved or brought to an acceptable level of risk:

(a) Compatibility of concepts (that reported electricity payments match well to what is on administrative files);

(b) Sufficient coverage (not all provinces are covered yet and the meter-to-dwelling integration may reduce the quantity available in certain areas);

(c) Data supply risk (requires systematic data supply);

(d) Cross-country uniformity (concepts are relatively simple in this case);

(e) Mode uniformity (analysis is on-going: respondents often provide rounded figures).

28. This work will help reduce the burden on respondents as this is a relatively difficult-to-answer question.

C. Commuting to work

1. Research to use spatial datasets and software for journey to work

29. One of the derived variables that the census disseminates is the distance from home to work. Roughly speaking, it is the straight-line ("as the crow flies") distance between the "blockface edge" of the home address and the blockface edge of the workplace address (for those with a usual place of work). This derived variable has been used in Statistics Canada's census dissemination for decades.

30. The major recognized limitation of this derived variable is that very few commuters take a straight line from home to work. Road networks and land features (rivers, mountains, etc.) all alter the trajectory of our journey to work. Therefore, this distance measure is for the most part underestimating the actual distance of most people's commute to work.

Figure III
Straight Line or Euclidean Distance
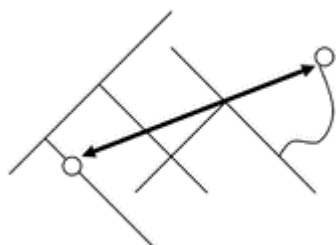
31. In the era of Google Maps and advanced geospatial datasets, it is hard to justify the continued use of a straight-line distance measure. Therefore, Statistics Canada has been exploring the use of geospatial datasets and tools to calculate the most probable home-to-work route of commuters ("network distance") in order to provide a much more realistic value of actual travelled distance from home to work.

Figure IV
Network Distance
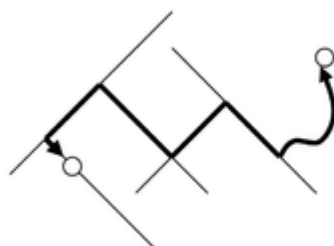
32.   Various geospatial datasets are being evaluated, including DMTI Route Logistics, the National Road Network File (NRN), open data from transit authorities, and Google Maps application programming interface (API). The tools used to assess these datasets include ArcGIS, Python, Google Maps API, Open Street Map and Azure Server.

2. Using latitudes and longitudes with census data

33.   The idea is to use the actual latitude and longitude of someone's geographic location (dwelling), rather than using the latitude and longitude associated with the Statistics Canada blockface edge associated with that location.

34.   In the context of the commute to work, using the actual latitude and longitude of the place of residence as well as the actual latitude and longitude of the place of work will provide a more realistic assessment of the commute to work, compared with using the latitude and longitude of the blockfaces of the residence and place of work. Blockfaces, particularly outside urban areas, do not need to respect road networks. For example, the blockface may run along a river edge, and the building in question is on the corner of that blockface and another, and at that point there is road access. In this scenario, the ability of the computer programs to determine the most logical route from A to B may be complicated by the fact that the latitude and longitude coordinates being (currently) used are on a river edge, not a road network.

35.   Essentially, it becomes harder for the computer programs to determine the most logical commuting route when the exact latitude and longitude are not used.

36.   This research is still in its early stages.

3. Limitations of using administrative data for commuting to work

37.   There are a few limitations to using this new approach to calculate distance from home to work. The first limitation is resource-based – both human and financial. Data need to be acquired and this creates new work for which people will need to be engaged.

38.   Second, route calculations are not 100 per cent accurate. Further restrictions for calculating routes are needed to reduce errors.

D.   Collectives

39.   A collective dwelling refers to a dwelling of a commercial, institutional, or communal nature. Examples of collective dwellings include lodging or rooming houses, hotels, motels, tourist establishments, nursing homes, hospitals, staff residences, military bases, work camps, jails and group homes.

40.   Statistics Canada employs a number of mechanisms to manage the response burden on Canadians by using existing administrative data sources, such as administrative records, to enumerate those living in institutional collective dwellings. These records have always been used. However, since 2016, Statistics Canada has been exploring new uses of administrative data for collectives including Canadian Forces bases, nursing homes and correctional facilities.

1. Canadian Forces bases

41. One of the areas for which administrative data are being explored is persons living in certain dwellings: Canadian Forces bases, (including other buildings on the base such as base hospitals), barracks, and guardhouses.

42. Data being received by Statistics Canada were investigated to see whether it would be possible to perform record integration to count individuals living on the base. However, early work led to a decision to enumerate military bases with field visits, as was done prior to the 2016 census, due to concerns about the quality of the records. These administrative records were not sufficiently detailed for the original purpose, but the data will be used to develop occupancy indicators for private dwellings in mail-out areas for the 2021 census.

2. Public guardians and trustees

43. Statistics Canada is also exploring whether it is possible to use data from public guardians and trustees to supplement data collection for some residents of collective dwellings including nursing homes, residences for senior citizens and residential care facilities. This type of information would be used to place the right people in the right location.

44. Currently, when a person's address points to the office of a public trustee, Statistics Canada is alerted to the fact that they may be in a collective. People in collectives seldom receive their mail directly and if someone's family is taking care of their affairs, there is a potential to erroneously record the senior with the family. So this administrative information may help for some cases, to record the senior in the right place.

45. The data on this are not currently available at Statistics Canada and require agreements with each province, so efforts to acquire the data and use them have been pushed back to be part of the work for the 2026 census.

3. Correctional facilities information

46. Statistics Canada is exploring using administrative files available for information on persons living in correctional facilities such as federal correctional facilities, provincial/territorial custodial facilities, and young offenders' facilities. The work is underway to explore the possibility of replacing field enumeration of these facilities for the 2021 census with administrative data. In the 2016 census, there were 813 collective dwellings of this type enumerated.

III. Conclusion

47. Administrative data offer an opportunity for Statistics Canada to improve the quality of census data, reduce response burden and expand census content. Nevertheless, there are challenges to overcome in order to successfully take advantage of administrative records for these purposes.

48. Among the key challenges is that of acquiring the data required. Often the administrative records are held by other levels of government, where agreements for transferring the information must be negotiated. Sometimes the quality of the administrative data are not sufficiently high for substitution.

49. Nevertheless, Statistics Canada has several examples of very successful data replacement, including income data and immigration information. For the 2021 census, the use of immigration data records for immigration information will be expanded, subject to final testing and analyses.

50. Administrative data sources also provide an opportunity for innovation, as evidenced by the research and development of new methodology for deriving journey-to-work content at Statistics Canada. Likewise, opportunities for improving information on those living in collectives are being pursued. However, data

acquisition and the quality of the administrative records are important considerations in the case of collectives.