Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Using data linkage to evaluate the consistency of place of residence between census data and tax data

Julien Bérard-Chagnon and Georgina House[1]

Abstract

Tax data are being used more and more to measure and analyze the population and its characteristics. One of the issues raised by the growing use of these type of data relates to the definition of the concept of place of residence. While the census uses the traditional concept of place of residence, tax data provide information based on the mailing address of tax filers. Using record linkage between the census, the National Household Survey and tax data from the T1 Family File, this study examines the consistency level of the place of residence of these two sources and its associated characteristics.

Key words: Census, tax data, place of residence, address, consistency.

1. Introduction

Tax data are being used more and more to measure and analyze the population and its characteristics. These data not only support different statistical processes, such as surveys, but are also used directly to develop a number of data and analytical products. While this approach comes with many benefits, it also poses a certain number of challenges. One of the main issues is the definition of the concept of place of residence. While several programs, including the census and population estimates, are based on the traditional concept of place of residence, tax data provide information on the mailing address of tax filers. Yet, place of residence is a fundamental element in examining populations. The majority of statistical indicators used to shed light on key socioeconomic issues rely on the ability of data sources to put people in the "right place" (National Research Council 2006). In this regard, conceptual differences between the census and tax data are likely to have a significant impact on the comparability of the two sources and, as a result, on the resulting products.

This study examines the consistency level of place of residence in these two sources and its associated characteristics using record linkage between the census, the National Household Survey (NHS) and tax data from the T1 Family File (T1FF). The next section discusses the concept of place of residence and the differences between the census and T1FF in this regard. Section 3 then introduces the data linkage used for this analysis. Section 4 presents the consistency level of the place of residence between the two datasets for different geographic levels. Finally, Section 5 examines the consistency level by various individual characteristics.

2. Concept of place of residence

Although it may appear simple at first glance, place of residence is a very complex concept. While most of the population is able to determine its place of residence with a high level of certainty, this information is more difficult to determine for some people. Children in shared custody and some students and workers regularly move back and forth between more than one place of residence, while the homeless, by definition, have no place of residence. Furthermore, some individuals may perceive their place of residence not as the location where they spend the most

[1]Julien Bérard-Chagnon, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (julien.berard-chagnon@canada.ca); Georgina House, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (georgina.house@canada.ca).

time, but where they maintain the strongest economic or social ties, such as young adults leaving the family home for the first time. Besides, problems in determining place of residence constitute one of the main sources of coverage error in the census (Statistics Canada 2015: 30).

2.1 Differences in the concept of place of residence between tax data and the census

There is no standard definition of place of residence. This concept changes from one source to another according to the main use of the data. These sometimes divergent definitions mean that the same individual may have a different place of residence from one source to another.

Canadian censuses use a legal concept of residence. Consequently, the address refers to the usual place of residence, which is defined as the dwelling in Canada where a person lives most of the time.[2] This approach is required for adequate planning of community services such as schools and public transit, allocating funds to the various levels of government and redistributing electoral districts. The census form includes rules for cases where the usual place of residence is more difficult to establish. For example, students who return to live with their parents during the summer must be enumerated in the family home even if they spend a significant—if not most—part of the year elsewhere.

Tax data provides information on the basis of tax filers'[3] mailing address. This address is required so that the Canada Revenue Agency (CRA) can contact the tax filer. The information is not intended to determine where the tax filer lives most of the time, but rather where they can easily be reached. Tax filers may well provide a mailing address that is not their usual place of residence, such as a very elderly person whose tax return is completed by a family member or an accountant.

These conceptual differences, combined with situations in which the place of residence may be more difficult to establish, can result in a portion of the population not having the same place of residence on the census and in tax data.

3. Data used

This study uses the data from a record linkage between the 2011 Census and tax data from the 2010 T1 Family File (T1FF). The T1FF is a file constructed by Statistics Canada that combines various tax sources, mainly the T1 individual tax returns and data from the Canada Child Tax Benefit (CCTB), to reproduce the Canadian population and families. This dataset was selected for this project because its coverage of the population is very high (95%) and it is often used by different analysts and researchers, especially to calculate official internal migration estimates. Data from the National Household Survey (NHS) were then added to the linkage to take advantage of the wealth of characteristics available in that dataset.

A deterministic approach was applied to the linkage using proven techniques. Individuals were linked in five successive waves based on their name, date of birth, sex and family information. Given the study's objectives, geographic information was not used as it would have biased the comparison; the linked persons would have been far more likely to have the same place of residence in the two sources.

The linked file contains just over 18 million people, or a linkage rate of 57.0% of the T1FF population. The NHS portion of the linkage contains about 4 million people. While the linkage contains slightly more children and slightly fewer people aged 65 and older, these differences are modest so that the linked data remain true to the T1FF and census data.

---

[2] The detailed definition of the concept of usual place of residence of the 2011 Census is available at http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/pop126-eng.cfm (page visited on November 5, 2015).

[3] Tax returns also include the province or territory of residence on December 31 of the taxation year and at the time of completing the return.

3.1 Definition of place of residence for this study

In this study, place of residence is defined based on the postal code. This decision stems from the fact that postal codes represent a very fine geographic unit close to the actual place of residence of individuals, primarily in urban areas, and because of the difficulties associated with harmonizing individuals' complete addresses.

The postal codes of the linked file were cleaned using a list of valid postal codes from May 2011. Based on this criterion, 95.9% of matched individuals had valid postal codes and thus could be used in this study. Place of residence was considered to be matched if the T1FF postal code was identical to that of the census.

3.2 File reference dates

A key element to remember in this analysis is that the reference date of the T1FF differs from that of the census. While the 2011 Census was administered on May 10, 2011, the 2010 tax returns were mostly completed in March or April 2011. Furthermore, the geographic information sent to Statistics Canada by the CRA to build the T1FF was the most current information available to the CRA as of December 31, 2011. For this reason, the data for the 2010 taxation year, collected in the spring of 2011, contain some geographic information from the end of 2011 if the tax filer changed his place of residence after completing his return, updated his information and the CRA incorporated that information in its databases. This difference may impact the comparison carried out here with respect to individuals who moved.

4. Postal code consistency level by geographic level

The table below shows the consistency level of place of residence between the census and the T1FF for different geographic levels.

Table 4.1-1
Postal code consistency level (in %) by different geographic levels

Geographic level	Consistency level
Postal code	92.9
Forward sortation area	95.5
Census subdivision	93.9
Census division	97.9
Census metropolitan area	98.7
Province or territory	99.7

Overall, 92.9% of matched individuals have the same postal code in the T1FF and the census. This rate climbs to 95.5% for forward sortation areas (FSA)[4] and to 93.9% for census subdivisions (CSD).[5] The slightly lower consistency level between CSDs and FSAs may be due to the fact that while the FSAs have a very fine geographic level in highly urban areas, they can be much larger areas than CSDs in the rest of the country. The consistency level continues to rise for the most aggregated geographic levels, approaching 98% for census divisions (CD)[6] and exceeding 99% at the provincial/territorial level.

These results tend to indicate that while the place of residence in the tax data may differ from the census, that difference is only perceptible at very fine geographic levels. However, despite the relatively high levels of consistency observed,

[4] An FSA corresponds to the first three characters of the postal code. In May 2011, Canada had 1,638 FSAs.
[5] A CSD is the general term for municipalities (as defined by provincial/territorial legislation) or areas treated as municipal equivalents for statistical purposes (e.g., Indian reserves, Indian settlements and unorganized territories).
[6] A CD is a group of neighbouring municipalities joined together for the purposes of regional planning and managing common services (such as police or ambulance services). These groups are established under the laws in effect in certain provinces of Canada.

two elements should not be overlooked. First, while 92.9% of matched persons have the same postal code, 7.1% of individuals in the linkage do not. If the linkage results can be applied to the entire population, close to 2.5 million people apparently did not give the same postal code on the census and in their tax data. This number is equivalent to the population of the Vancouver census metropolitan area (CMA),[7] the third most populous CMA in Canada.

Second, there are sometimes significant regional disparities in terms of the consistency level of postal codes. In particular, consistency is weaker in the territories where it reaches a low of 85.1% in the Northwest Territories. This result means that fewer than 9 out of 10 residents in this territory have the same postal code in the T1FF as on the census. Furthermore, at local levels, the consistency level generally tends to decline in urban cores and be higher in peripheral areas. The population living in downtown Montréal has a postal code consistency level below 85%, the lowest level of the entire Montréal region. Inversely, the population living in several suburbs and in certain areas in the West Island regularly have a consistency level above 95%. These results are essentially due to the combined impact of regional differences in population characteristics and the links between those characteristics and postal code consistency levels.

5. Postal code consistency level by certain characteristics

The following table presents the consistency level of postal codes by various characteristics.[8]

Table 5-1
Postal code consistency level (in %) by various characteristics

Characteristics	Consistency level
Total	92.9
Age group	
0 to 9 years	91.5
10 to 19 years	93.6
20 to 24 years	91.0
25 to 29 years	89.2
30 to 39 years	92.0
40 to 49 years	94.7
50 to 59 years	95.4
60 to 69 years	94.9
70 to 79 years	93.1
80 years and older	79.9
Moved in the last year*	
Non-migrant	95.4
Migrant	83.0
Lives in a collective dwelling	
No	93.4
Yes	31.4
On-reserve household*	
No	94.2
Yes	87.6
Province/territory of residence and work † *	
Same province/territory	94.5
Other province/territory	89.4
Spouse present in household*	
Spouse present	95.0
Spouse absent	89.0
Linkage wave	
1) Exact match	94.6
2) Less restriction on the name	93.3
3) Mix match	93.1
4) At least three members of the household with same sex and birth date	92.3
5) Single persons added	84.8

* Characteristics from matched persons who responded to the NHS. Because of the study objectives and the linkage rate, the results presented here that are drawn from the survey are not weighted.
† Persons living in the Ottawa–Gatineau CMA and in the census agglomerations of Campbellton, Hawkesbury and Lloydminster were not included in this analysis because their metropolitan area overlaps two provinces.

[7] A CMA is an area consisting of one or more adjacent municipalities situated around a major urban core. A CMA must have a total population of at least 100,000 of which 50,000 or more must live in the core.
[8] While the results in this study are solely descriptive, the links between these characteristics and the postal code consistency level remain statistically significant when the impact of various factors at taken into account simultaneously using a logistic regression model.

The data in the table show that a number of characteristics are associated with lower consistency levels between the census and T1FF postal codes. An examination of the postal code consistency based on age reveals two things. First, the consistency level dips moderately beginning at age 20, falling below 90%, and then starts to recover around age 30. This result is explained in part by the greater likelihood to move at these ages. In fact, this stage of the life cycle is generally marked by a number of moves motivated by leaving the family home, attending postsecondary institutions, entering the workforce or purchasing a first property. As a result, the likelihood of moving peaks at these ages (Dion and Coulombe 2008). A move in the year preceding the NHS is also associated with a lower postal code consistency level. While individuals who have not changed addresses show a consistency level of more than 95%, this rate falls below 85% for those who have moved.

Consequently, a number of characteristics associated with mobility, such as immigrant status or being a tenant, are also linked to a slightly lower consistency level. While this situation may be due to the fact that individuals who move are more likely to continue to maintain ties, at least temporarily, with more than one place of residence, it can clearly also result in large part from the discrepancy in the reference dates of the two sources. It is noteworthy that even after taking into account the fact of having moved, young adults still post a slightly lower consistency level, suggesting that factors other than mobility may be the reason for this situation.

Second, the consistency of postal codes declines significantly among the very elderly. It falls below the threshold of 80% beginning at age 80 years and reaches almost 60% among those 90 or older. This decline is explained by the strong propensity of seniors to live in collective dwellings. The postal code consistency level is especially low for individuals living in a collective dwelling (31.4%). If we only look at persons living in private dwellings, the postal code consistency level of individuals aged 80 and older remains at around 90%. The very elderly are also more likely to have their tax return prepared by an accountant or a family member. This factor might be associated with a lower postal code consistency level due to the fact that the postal code on the return could be that of the person preparing the return and not the tax filer.

Certain characteristics associated with situations where place of residence is more difficult to determine also correlate to slightly lower consistency levels. Thus, individuals who do not work in their province or territory of residence and those who do not live with their spouse show consistency levels below 90%. In addition to being more mobile than the national average, these individuals are also particularly likely to maintain ties with more than one place of residence so that they might enter different postal codes on the census and in tax data. Individuals living on an Indian reserve also have a somewhat lower consistency level (87.6%). The economic, social and political dynamics of these areas are sometimes quite different from those in the rest of the country.

Lastly, the consistency of postal codes also follows a gradient based on the matching wave. Individuals linked in the fifth wave post a particularly low level of consistency (84.8%). The approach used to construct the linkage means that links created in the final waves are considered slightly less robust so that the chances of a false match are a bit higher. Clearly, in this situation, postal codes are much less likely to match. Moreover, an examination of the individuals matched in this wave reveal that they are more mobile and more likely to live in a collective dwelling, two characteristics also associated with lower postal code consistency levels.

6. Conclusion

The place of residence concept is very complex. Conceptual differences between different data banks can impact the coherence of statistical products generated from them. This study examined the consistency of the place of residence between T1FF tax data and the census using data linkage.

The analysis revealed that the consistency level of place of residence is relatively high. Indeed, 92.9% of individuals matched had the same postal code in the census and the T1FF. This proportion climbs to more than 99% when considering the most aggregated geographic levels. This study also brought to light certain, sometimes major, variations in the consistency levels of different segments of the population. Young adults, the very elderly, highly mobile individuals and persons living in a collective dwelling are particularly likely to post lower consistency levels.

In general, in the context of a growing use of administrative data and the diversification of social trajectories, data comparability exercises such as this one will become increasingly important not only to better understand the files available but also to better measure the demographic dynamics of the Canadian population.

References

Dion, P. et S. Coulombe (2008), « Portrait of the mobility of Canadians in 2006: Trajectories and characteristics of migrants », *Report on the Demographic Situation in Canada 2005 and 2006*, Statistics Canada, 91-209-X, pp. 83-134.

National Research Council (2006), *Once, Only Once, and in the Right Place: Residence Rules in the Decennial Census*. Washington, DC: The National Academies Press, 355 p.

Statistics Canada (2015), *Census Technical Report: Coverage. Census of Population, 2011*, 98-303-X, 163 p.