Issues regarding the use of the Canadian census sample of 1852 for data linkage

Catalina Torres
Programme de recherche en démographie historique (PRDH), Université de Montréal
catalina.torres@umontreal.ca

Abstract
This paper discusses some issues that should be considered when using the Canadian census sample of 1852 in data linkage projects.

Introduction
Between 2004 and 2006, the PRDH created a 20% sample of the first nominal census of Canada: the census of 1852. Despite critiques of this census, the 1852 enumeration offers unique opportunities to researchers wishing to study the Canadian population of the mid-nineteenth century, particularly the rural population (Dillon and Joubert, 2012). Both the census sample of 1852 and the 100% database of the 1881 Canadian census constitute rich sources of information about the Canadian population of the mid- and late-nineteenth century. For example, both sources contain valuable socioeconomic variables and provide information at the individual level, making these data suitable for record linkage. If the intervening time span of 30 years seems very long, it is perfectly suitable for the study of certain transitions.

In the course of our work in the PRDH, we have noticed several issues which should be considered when using the Canadian 20% census sample of 1852 for linkage purposes. Some of these are: 1) the representativeness of the urban data, 2) the interpretation of the inscriptions of some fields that are usually used on linkage protocols, such as the place of birth, and 3) the homogeneity of French last names.

1. The missing urban data
One very important issue regarding the Canadian census of 1852 is the absence of one third of the records: the census manuscripts covering 34% of the population disappeared before being microfilmed (Dillon and Joubert, 2012). Most of these missing manuscripts contained the records of the urban population: in Upper Canada (UC), the records of the cities of Toronto, Kingston, London and the big district of Simcoe are missing; in Lower Canada (LC), the records of Montreal are lost, except for those of the neighborhood of St. Louis.

Fortunately, some urban data are available— from Bytown and Hamilton in UC and from Québec City in LC. By attributing weights to these records—a weight variable is available on the 20% sample—some aspects of the urban population of mid-nineteenth-century Canada can be analyzed (Dillon and Joubert, 2012).

Despite the availability of some urban data, and even if the majority of the Canadian population in the mid-nineteenth century was rural, the missing records are still problematic for any linkage procedure at the individual level, since most of the people living in the big cities in 1852 will be excluded. Attributing weights is certainly not always a solution to this problem, since the population living in big cities such as Toronto or Montreal very likely differed in several aspects from that living in smaller cities such as Hamilton or Québec City. Thus, regarding record linkage with the 1852 census, the absence of most of the urban data necessitates caution, both in our methods and in our interpretations: the urban individuals available for linkage will represent only certain selected cities, leaving the biggest cities—Toronto and Montreal—underrepresented.

1 The PRDH has lengthy experience with the record linkage of Québec parish registers, and more recently undertook linkage of a sample of the 1871 Canadian census to the 1881 census; we are currently involved in the Mining Microdata project to link the 1852 and 1881 censuses. Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850-1911, Digging Into Data Challenge, ESRC/NSF/CRSH, team (Evan Roberts, leader), 2012-2014.

2 One example is intergenerational social mobility: by linking the sons across a time span of 30 years intergenerational social mobility can be measured by comparing the occupation of fathers and sons at similar points in their lives, thus avoiding the problems that arise when this type of mobility is measured by the marriage records only (Van Poppel, De Jong, et Liefbroer, 1998).
Thus, researchers using the Canadian census sample of 1852 in record linkage projects should consider whether the missing urban data constitutes a problem to the aims of their specific study or project. For example, the missing urban data should not be problematic in the analysis of the rural exodus, which could be studied by linking the census sample of 1852 with the complete enumeration of 1881. It should be kept in mind that the urban population of Canada in 1852 constituted only 10% of the population, and many of the missing manuscripts are distributed across communities of varying sizes.

2. “B” for “Bas” or for “British”?
We have also discovered some ambiguities regarding the codes attributed to the place of birth inside Canada in the census sample of 1852. For example, birthplace strings such as “Canada b” and “b Canada” were initially coded as Québec (known prior to 1841 as Bas Canada or Lower Canada), as the “b” was probably associated with “Bas”. However, we have considered the possibility that the letter “b” may not always mean “Bas”, e.g. when the enumerator was Anglophone.

The analysis of previous work done by the PRDH as well as the visual inspection and analysis of census manuscripts and data were useful to solve this problem. First, a new variable “B” was created in order to identify the cases where the place of birth is accompanied by the letter “b”, for example “b Canada”, “Canadian b”, but also more precise places such as “Montreal b”, Beaumont b”, etc. Second, a new code was created within the variable birthplace. This new code was attributed to those cases in which the place of birth 1) is accompanied by the letter “b” but lacks precision (for example “Canada b”) or 2) refers to the English origin of the individual.

A similar solution had been applied to cases where, instead of a “b”, an “f” accompanied the place of birth. For certain cases in which the response was “Canada f”, a specific code was created indicating “Canada French (province not specified)”; this value includes some mentions relative to the cultural group of the individual, such as “french canadian”. As to the visual inspection of the data and the manuscripts, most of the individuals (≈97%) to whom the new code concerning the letter “b” was attributed have an English last name. In addition, the content of several census pages suggests that the enumerators did sometimes fill the column place of birth with mentions relative to the cultural origin of the individual (for example “Canadian f” and “Canadian b” are frequently found in the same page, and they generally correspond to French and to English last names respectively).

In the entire 20% census sample of 1852, the proportion of individuals identified with “f” or with “b” as well as with the corresponding birthplace code indicating “Canada F” or “Canada B” (province not specified) is about 10% and 1% respectively. We believe that these new codes better document the mentions inscribed on the census manuscripts, indicating the cases in which a specific province of birth cannot be attributed, and the cases in which “f” and “b” suggest the cultural origin of an individual rather than a place of birth.

The variables/codes discussed in this section, which are the result of recent work by the PRDH, have not been used in a linkage protocol yet. However, they may be useful in the stage of manual verification. For example, regarding the cases in which the birthplace is not specified (e.g. “Canada” in one census and “Québec” in the other), the consideration of additional variables, such as the cultural origin of the individual, during the manual verification phase may help to assess the validity of a link.

3. The homogeneity of French last names
A more general observation which applies to all the 19th-century Canadian censuses is the homogeneity of the last names of French origin. In Canada, the stock of French surnames is indeed much more limited than the stock of English-origin surnames, since the French-Canadian population is descended from basically 10,000 French immigrants who arrived in Québec before the 1760s. In contrast, the regular immigration to Canada of people from the British Isles during the 19th century nourished the pool of English last names (Charbonneau et al., 2000). According to Dillon (2002), who examined the distribution of last names in the complete enumeration of 1881, the ratio of French and of English Canadians to a single last name was of 46:1 and of 22:1 respectively, demonstrating that more French-Canadians shared the same surname than English-Canadians did.
Regarding data linking, this homogeneity implies that focusing on uncommon names is inappropriate as linking strategy for the French Canadian population (Dillon, 2002). To assess the validity of a link, a set of additional variables should then be considered, such as the approximate age expected at one point in time, the gender, the (logical transition of the) marital status, the place of birth, the cultural origin, the religion, etc.

Since the last name is one of the variables commonly included in linking protocols, the homogeneity of French last names raises concern about the representativeness of the population that would be linked using the census sample of 1852. However, as is frequently remarked on studies based on linked records, common names may be under-represented, but that does not necessarily introduce bias into the linked sample (Dillon, 2002).

Conclusion
This paper has briefly discussed the three following issues regarding the use of the Canadian census sample of 1852 in the context of projects that involve data linking: first, the absence of the majority of the urban records; second, the modifications to the variable place of birth in order to take into account the differences introduced by the Canadian cultural environment of the mid-nineteenth century; and third, the homogeneity of the last names of French origin. Our awareness of these issues is important to avoid bias in studies willing to use this source for linkage purposes, among others. As a final consideration, one possibility to verify the validity of the links is to use the marriage registers (BALSAC), which are available for the catholic population in Québec until the early 20th century. Thus, the information about the spouse and other members of the family could be used to assess the validity of a link3.

References

---

3 Such verification will be done soon by the PRDH using the linked data from the census sample of 1852 and the complete enumeration of 1881. The registers of the catholic marriages that took place in Québec the period between those two censuses will be used in order to examine the accuracy of the links.