The linkage of micro census data and vital records: an assessment of results on Quebec historical censuses (1852-1911)

Hélène Vézina
Projet BALSAC, Université du Québec à Chicoutimi

Marc St-Hilaire
Centre interuniversitaire d'études québécoises, Université Laval

Jean-Sébastien Bournival
Université Laval

Paper presented in the session Challenges, Methods and Solutions in Record Linkage of Micro-data at 40th Annual meeting of the Social Science History Association, 14 November 2015

A preliminary version of this paper was presented in the session: "Using micro census data: recent developments and new perspectives" at the First Conference of the European Society of Historical Demography (ESHD) in September 2014

ABSTRACT

We recently completed the development of a linkage program to match microdata from the Canadian censuses to those from Quebec civil records found in the BALSAC population database. Procedures underlying this software are modeled on the basic principles used at BALSAC for the linkage of civil records. Several adjustments have been made however to take into account differences in the structures of the two types of data. Ongoing work covers three Quebec regions and two cities for seven modern nominal censuses (1852 to 1911).

In this paper, we describe the program and analyse the results of linkage operations performed in the Saguenay region and the city of Trois-Rivières. In order to provide a critical assessment of these results, we investigate household and individual characteristics that could impact on linkage rates. We also illustrate the potential of the program by comparing linkage results among men and women. Lastly, we discuss the possibilities and limits involved in our approach

The linkage program is at the heart of the construction of the Integrated Infrastructure of the Quebec Population Historical Microdata which will integrate all available historical microdata on the Quebec population, dating back to the beginning of European settlement, into a set of relational databases. The interconnection between civil records and the censuses, along with linkages across the censuses themselves, will substantially broaden and enrich the avenues of research in both the social and the biological sciences. Once linkage is completed, it will be possible to carry out studies based on individual trajectories within families, households and communities from intra- and inter-generational perspectives.

Introduction

The Quebec province in Canada has the advantage of an exceptional and remarkably well-preserved documentary resource with the wealth of birth, marriage and death certificates recorded by the Catholic Church since the early days of French settlement in the 17th century. More than 40 years ago, two major projects, the Programme de recherche en démographie historique (PRDH) and BALSAC, initiated the digital transcription and linkage of these parish records. More recently, a good share of Quebec data from the Canadian historical censuses prior to 1921 have also been digitized and formatted in databases by the Centre interuniversitaire d'études québécoises (CIEQ) and by the PRDH.

These two sources of data are fundamental for the study of two major categories of social facts. Information from vital events, once linked to create family files, fosters research on family and kinship, such as reproduction, union formation, family strategies, migration and intergenerational transmission In turn, nominal censuses make possible studies related to the household (size and composition, roles of members, residential patterns, occupational structures, income and education) as well as micro-scale economic strategies.

With continuous data found in civil registration, it is possible to conduct longitudinal studies using family reconstitution techniques while sophisticated multivariate analytical tools have been developed to investigate census data which is by nature cross-sectional. Each source also has its limitations: vital registration provides none or very little indication on the socioeconomic conditions of individuals and families and none on living arrangements and neighborhood; censuses do not provide information on the role of extended kinship in economic and residential strategies.

In order to overcome these limitations and build on the strengths of the two types of data to expand research possibilities, we initiated in 2010 the development of a linkage program. The program relies on systematic and automated tools and procedures to match microdata from the Canadian censuses to those from Quebec civil records and to link census data together. We aim to develop a research infrastructure that will take advantage both of the power of family and genealogical files based on civil registration and on the wealth of census data and to open up new avenues of research for the study of historical populations.

In this paper, we describe the two components of the program and analyse the results of linkage operations performed in the Saguenay region and the city of Trois-Rivières. In order to provide a critical assessment of these results, we investigate household and individual characteristics that could impact on linkage rates and affect the representativeness of the linked population. We also illustrate the potential of the program by comparing linkage results among men and women. Lastly, we discuss the possibilities and limits involved in our approach.

Population reconstruction projects in Quebec

The Registre de la population du Québec ancien (RPQA) was created in 1966 by the Programme de recherches en démographie historique at the Université de Montréal. Its construction was inspired by the techniques of family reconstitution developed by the French demographer Louis

Henry. It contains the longitudinal linkage of the entire Catholic population of Québec from the beginning of French settlement in Canada in 1608 (first record in 1621) to 1799 comprising 700,000 birth, marriage and death records.

In 1972, another population database was initiated at the Université du Québec à Chicoutimi. In the first phase, family reconstitution was performed on the population of the Saguenay-Lac-St-Jean, a region located 200 kilometers north of Quebec city (see Figure 1), from the beginning of French Canadian settlement in 1838 to 1971 relying on the 660,000 birth, marriage and death records. Since 1989, a second phase has been ongoing to perform the digital transcription and linkage of marriage records for the whole province of Quebec. The work is now completed up to 1965 and the database contains 2.2 million linked marriage records allowing for the automatic construction of genealogies.

The complete census of Canada for 1881 as well as samples for all other Canadian historical censuses (from 1852 up to 1911 with the exception of 1861) are formatted in databases and publicly available. In Quebec, more specifically, complete count for the cities of Quebec and Trois-Rivières for seven decennial censuses (1852-1911) and partial count samples for many Quebec regions have been digitized in the course of previous projects[1].

A few attempts to link Quebec civil and census records were previously carried outr relying on manual work performed on small datasets (Charbonneau *et al*., 1970; Gauvreau, 1991; Gossage 1999). In some instances marriages records were used to facilitate linkage across censuses (St-Hilaire and Marcoux 2006; 2014; St-Hilaire, 2009; Olson and Thornton 2011). In 2010, members of our group initiated a pilot project to evaluate the feasibility of designing a program to link census data to marriage records found in the BALSAC database (Gauvreau *et al*. 2010; St-Hilaire and Vézina, 2010; Vézina and St-Hilaire, 2011). The development of this linkage program is now at the heart of the construction of the Integrated Infrastructure of the Quebec Population Historical Microdata (IMPQ) which was financed in 2013 by the Canadian Foundation for Innovation.

The IMPQ project

The IMPQ will integrate available historical microdata on the Quebec population dating back to the beginning of European settlement, into a set of relational databases. The objective is to preserve, highlight and develop this major historical and scientific heritage. The creation of the infrastructure will also facilitate integration of new data and development of new tools for linkage and analyses as well as promote training in relevant disciplines and collaborations both at the national and international levels (comparative research).

The construction of the infrastructure includes three components. First we plan a full fusion of the BALSAC and PRDH databases enabling simultaneous and continuous updating of the longitudinal data as well as integration and linkage of new vital records. The second part

[1] The 1881 census is available in the North Atlantic Population Project (NAPP) website (https://www.nappdata.org/napp/). Data on Quebec City is available in the Population et histoire sociale de la ville de Québec project website (http://www.phsvq.cieq.ulaval.ca/index.php?p=accueil). The other datasets were constructed in the course of research projects and are not publicly available at this moment.

focuses on the harmonization of existing census data series and expansion of the geographical coverage to include two urban environments (Quebec City and Trois-Rivières) and three regions mixing rural and urban environments (Gaspésie, Côte-Nord and Saguenay) (see Figure 1). In the third component of the project, households and individuals from these areas are gradually linked to the BALSAC database and across censuses. The linkage program we present here is thus at the heart of the construction of the IMPQ.

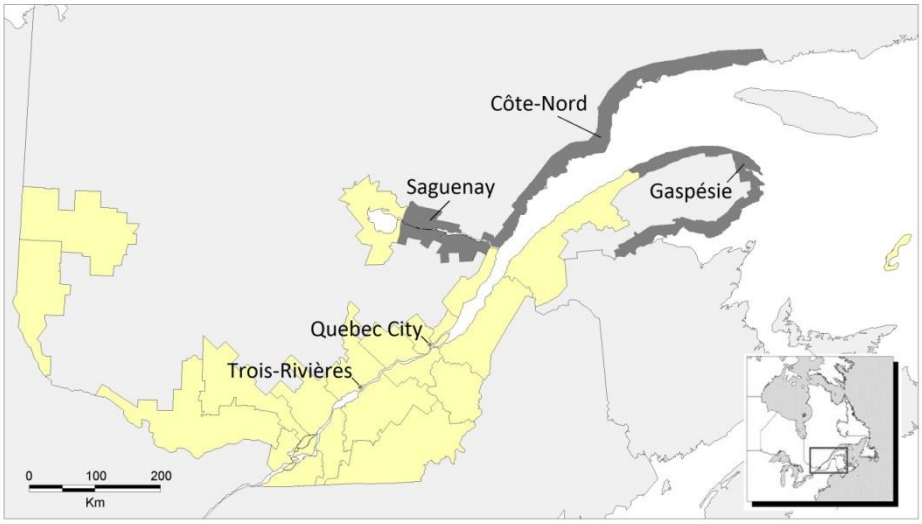

Figure 1: Census data to be integrated in the IMPQ

The development of a program for the linkage of civil records and census data

The BALSAC linkage system of vital records as a starting point[2]

From the start, we made the decision to use an approach based on the linkage procedures developed for the construction of the BALSAC database. The challenge was to adapt them to take into account the differences in the structures of the two types of data. We will first present briefly the BALSAC linkage system.

The BALSAC family reconstitution system, which processes vital data records, aims at grouping in the same file all mentions appearing in the records and referring to the same couple. Thus, the basic unit of information for linkage is the "couple mention" which contains four nominative elements namely the husband's name and surname and the wife's name and surname. All nominative information is processed to eliminate superficial name variations using FONEM, an automatic phonetization program (Bouchard *et al*., 1981a; Bouchard and Bouchard, 1981). As shown in Figure 2, FONEM can significantly simplify the name variations to be linked.

[2] This section of the paper will not be presented at the SSHA conference.

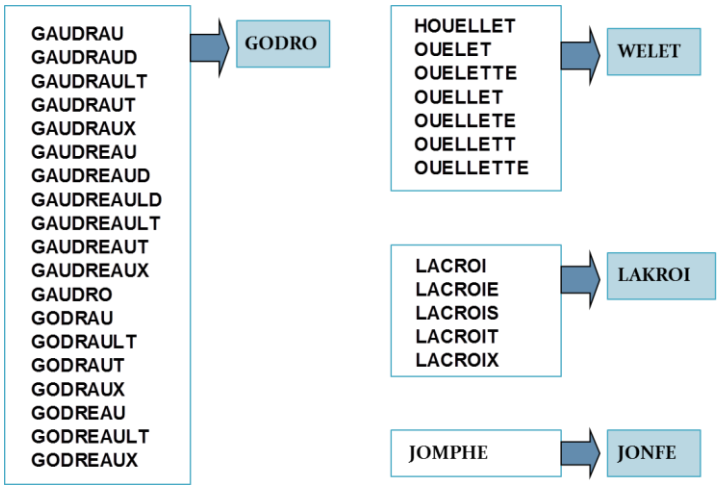

Figure 2: Examples of automatic phonetization with the FONEM program

Candidate couple mentions in BALSAC are created on the basis of at least two common nominative elements with the couple to be linked. These mentions are compared using three programs designed to detect and measure degrees and forms of similarity between two last names or two first names. ISG calculates a score based on the degree of similarity between nominative elements based on the position of same letters in names. INCL deals with truncated names by detecting suffix and prefix in names and deciding if one can be treated as being included in the other. ELM processes situation of multiple names and surnames and decide if two entities car be treated as equivalent or not. The decision making process relies almost exclusively on nominative information contained in the records but the coherence of dates in the family structure is also taken into account (Bouchard *et al.*, 1985).

The linkage operation consists of four different phases processing first the easiest links and moving progressively to the most difficult. The two first phases are performed immediately after the entry of a marriage record. The first one is entirely automatic: a couple is linked to an already existing family files only if all nominative information is perfectly identical and if there is only one potential candidate for the linkage. In the second phase, when the program has been unable to make the appropriate linkage within the established safety range**,** it provides a list of potential candidates for linkage and the entry clerk makes a selection if an unequivocal choice can be made. The third phase is performed at a later stage. It is also based on a list of candidates provided by the program and is operated by experienced staff who might use complementary sources such as genealogical repertories or websites to support their decision. Lastly, after these three stages, there remains a share of unsolved linkages which will eventually be submitted to a more in-depth investigation.

Linkage operations lead to the construction of a family file that includes all the records pertaining to a unique couple (their own marriage, the remarriage of a surviving spouse when existing, the marriage of their children[3]) (St-Hilaire, 1990). All links submitted to automatic

[3] For Saguenay, family files also include spouses' death records, children's birth as well single children death record.

linkage and those carried out at the stage of the computer-assisted manual linkage are immediately validated by automatic routines of coherences (for example: acceptable timespan between two events, chronological sequence of events, reported or calculated age, duplicate events or names)(Bouchard *et al*., 1981b).

Adapting the BALSAC method for census and civil records linkage

As mentioned in the previous section, the development of the program for the linkage of civil records and census data rests on the basic principles used at BALSAC for the linkage of civil records with nominative elements at the heart of the process. Essentially, the method remains the same and linkage is done using matching programs as well as comparison and decision algorithms developed at BALSAC. However, several adjustments were made to take into account differences in structure and content between the two types of data. Hence, before describing the method of linkage itself, the main issues raised by the differences in the two data sources and an overview of modifications brought to the BALSAC system are presented.

First, the structure of the basic grouping units (family vs. household) is different. The BALSAC family structure corresponds to the model of a "nuclear family" based on information found in civil registration events. Census household composition is determined by a survey at a specific point in time and is based on a residential criterion (housing). Several scenarios of household structure are possible: one or many single individuals, nuclear family (partial or complete), blended family, intergenerational family, etc. That being said, the similarities between the two types of data are real, since the household as a grouping unit often corresponds to a structure, if not identical, at least similar to that of a nuclear family.

Second, the quality of nominative information is not the same in both sets of data. Civil records are legal documents and priests, pastors, and other civil officers usually kept them with great care. They are very consistent trough time since there was little change in prescribed rules to record information (Bouchard and Larose, 1976). The context of census data collection is very different. It is a massive and complex operation aimed primarily at ensuring a fair representation at the Parliament but also used to follow the development of the nation. It is conducted in a short period of time, at ten years intervals by people who were not always well-trained and sensitive to the cultural context of the people they enumerated (Bellavance *et al*., 2007). These factors, combined to the fact that the census nominative information has no legal value, impact directly on the quality of the nominative data (usual instead of legal first and last names, nicknames, abbreviations, phonetic alterations, etc.).

Last but not least, whereas in the BALSAC database, first and last name are found for both spouses, in census data, the maiden name of the wife is almost always missing. As the whole linkage system relies on a couple mention comprising four distinct nominative elements, this represented an important issue for the development of our program.

Given these differences, some modifications were brought to the linkage program and new elements were introduced. One of the main changes in the method consists of a prior shaping of census data in order to extract nuclear families comparable to BALSAC families. To be

considered for linkage, a household must contain at least two members forming a familial unit (husband and wife or widow-er with a child) since a minimum of three nominative elements is needed.

To overcome the absence of the wife's maiden name in the households, it was decided to generate a new nominative unit of comparison (NUC) including the three other nominative elements available (father's last name, father's first name and first name of the mother) to which is added the first name of a child. It is then possible to compose as many NUCs as there are children mentioned. For each set of data, the program creates a table of nominative mentions whose number per household or per family varies according to the number of children.

Because of the changes brought in the composition of NUCs, it was necessary to relax sorting and matching criteria to bring more potential candidates in the phase of nominative pairing (see below). The adapted method of matching requires a single nominative element out of four to be perfectly identical between a NUC derived from the census and the one obtained from BALSAC. This also permits to overcome the difficulty to harmonize, standardize and compare nominative data from two different sources (civil registration and census). The relaxation of matching rules is however offset by the addition of a rating system of candidates which will be reported in the next section.

Another important change is the fact that elements of comparison other than nominative information are taken into account for linkage. Because of the variability in nominative information, it seemed important to involve other attributes in the process of comparison to support situations where the nominative context would be inappropriate. Thus, other elements found in the two sources were added for comparison, namely: residential data, occupations and information allowing for verification of concordance of dates and ages. These elements are not considered in the selection of household and family candidates, but are used in the comparison process and score assignment which quantifies the degree of similarity between the two sources. Hence, while linkage of civil records rests almost exclusively on nominative information, linkage of civil and census data relies on the comparison of both individual (nominative data) and contextual (family/household characteristics) elements.

Lastly, after performing some tests, we made the decision to exclude automatic linkage and to rely entirely on manual computer-assisted linkage. This means that all linkage decisions are taken by research assistants.

Linking census data to BALSAC families

The first module of the program is used for linking census data to BALSAC and involves five steps described on Figure 3. In the two first steps data from the census and from BALSAC are transformed into a common structure: the nominative unit of comparison (NUC). In step 3, the NUCs are sorted and paired to generate a pool of potential candidate families in BALSAC. In step 4, each candidate receives a score based on selected matching criteria. Lastly, in step 5, a

final decision is made to proceed with linkage or not. When a household is linked, the program allows for linkage of individuals belonging to this household.

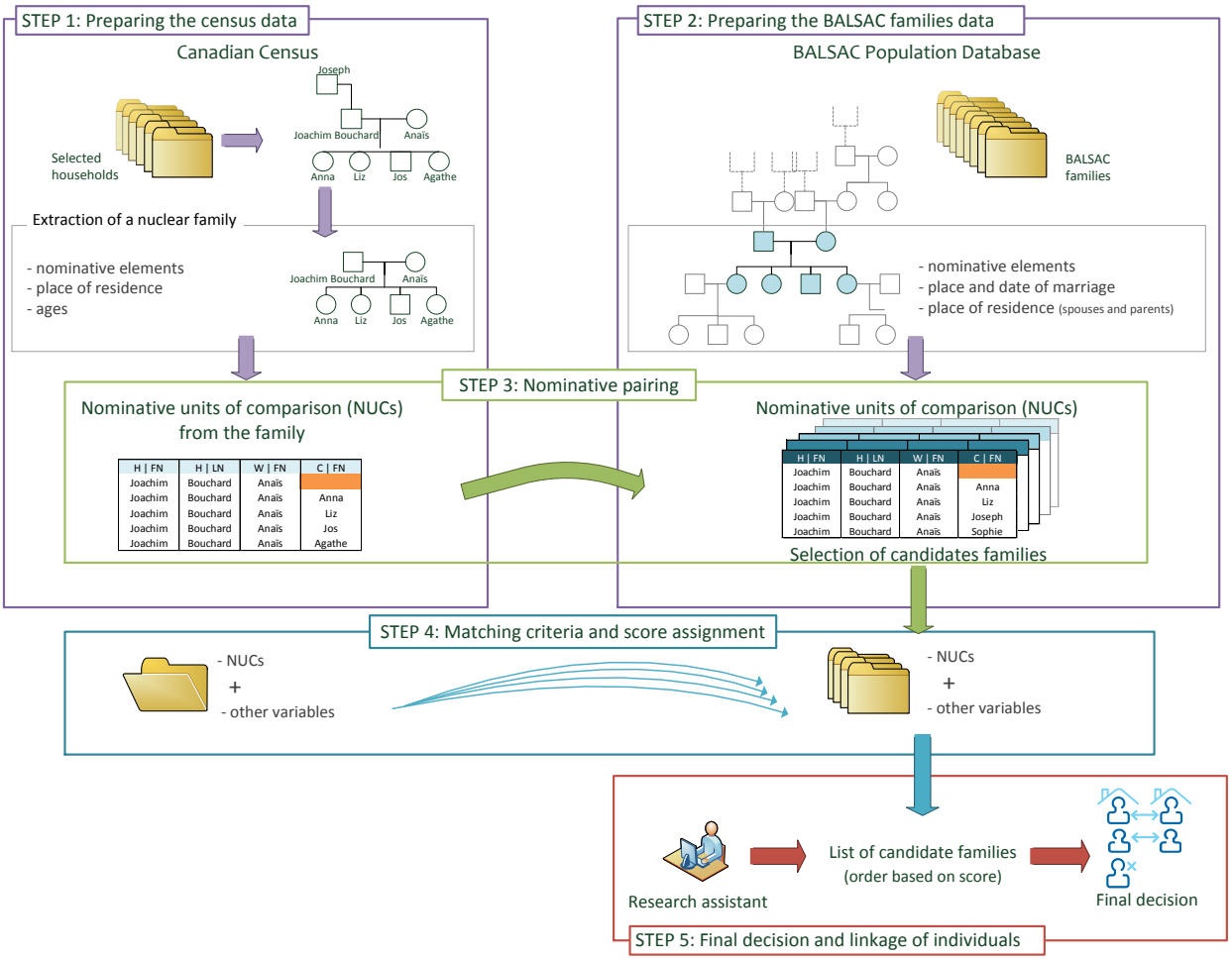

Figure 3: The linkage of census data to BALSAC families

Step 1. Preparing the census data: In order to form the NUCs, nuclear families must first be identified and extracted from census households. More than one family can be extracted from a household. To compensate for the absence of the wife's maiden name in census data, the first name of each child is added to the NUC. Hence, there will be as many NUCs for a census family as there are children. The three-element combination containing the first and last names of the husband and the first name of the wife is also used and it is the only combination possible when a couple has no children recorded. Other members of the household are regarded as children of this couple if a) their surname is the same as the presumed father; b) their age is compatible with the age of the mother (minimum age difference of fifteen years); and c) they are single. In addition to the nominative data, other variables that will be used to link records and to calculate a score are kept (place of residence, ages).

Step 2. Preparing the BALSAC families data: Each potential candidate family from BALSAC is also organized in NUCs to provide a pool of candidates for each census family. First, a subset of families is extracted from the BALSAC database. These families are selected within a certain

time interval based on the likelihood of having been enumerated at a given census. First and last name of husband and wife, name of children as well as other variables that will be used for linkage (place of residence, ages and date of marriage) are retrieved.

Step 3. Nominative pairing: Based on the programs of comparison of nominative similarity developed at BALSAC and described above, the application performs various sorting operations aiming at pairing each census family to candidate BALSAC families. The pairing criterion is that both NUCs (one from the census and the other from BALSAC) must contain at least one identical nominative element. A list of potential candidates for linkage with a specific census family is thus selected from the BALSAC families.

Step 4. Matching criteria and score assignment: Each of the BALSAC families selected in the pairing process is submitted to the comparison component of the program and a score is attributed to each potential match based on the degree of similarity with the census family being processed. While nominative data found in the NUCs still provides the most important linkage criterion, concordance of dates and ages as well as place of residence are also used in the comparison process. The goal is to take into account and make optimal use of the information available about the families in each dataset. Every element of the comparison receives a score weighted by its importance (for instance, logical concordance of dates is more important than similarity of children's first names) and the sum of these scores provides the total score.

Step 5. Final decision and linkage of individuals: At this stage, the candidate BALSAC families are ordered according to their final score and manual work is involved to proceed with the linkage decision. Based on the work presented in this paper, we have estimated that depending on the region and on the census year, between 72% and 92% of the created links involve the candidate with the highest score. Moreover, a final decision is taken to accept a candidate or abort the linkage attempt within two minutes on average. When a household is linked, the research assistant can also link individuals based on name, age and date consistency.

Linking families and individuals across censuses

The second module of the program allows for linkage across censuses. In the work we are currently conducting for the IMPQ project, linkage between two censuses takes place after both censuses have been linked to BALSAC using the first module. The whole linkage process is similar to the one described in Figure 3. In both censuses, nuclear families are extracted from household and NUCs are produced. Nominative pairing based on the NUCs lead to the selection of candidates from Census B for linkage to a specific family in Census A. Nominative and contextual information are used for score assignment and the final decision is taken by the research assistant.

There is however a prior selection in Census A of households depending if they have been linked to a BALSAC family or not. This selection process is displayed in Figure 4. It shows that in the preparation of data, some households will be processed and others will not based on the criteria described previously (at least two people in the household and the head being married

or widowed). Among processed households, some of them will then have been linked to BALSAC while for others linkage will not have been successful. Direct linkage to Census B will be attempted on all households from Census A which have not been linked to a BALSAC family. Moreover, among linked households from Census A, some will be connected to a household in Census B through their linkage to the same BALSAC family but others will not. Direct linkage to Census B will also be initiated on this latter set of households.

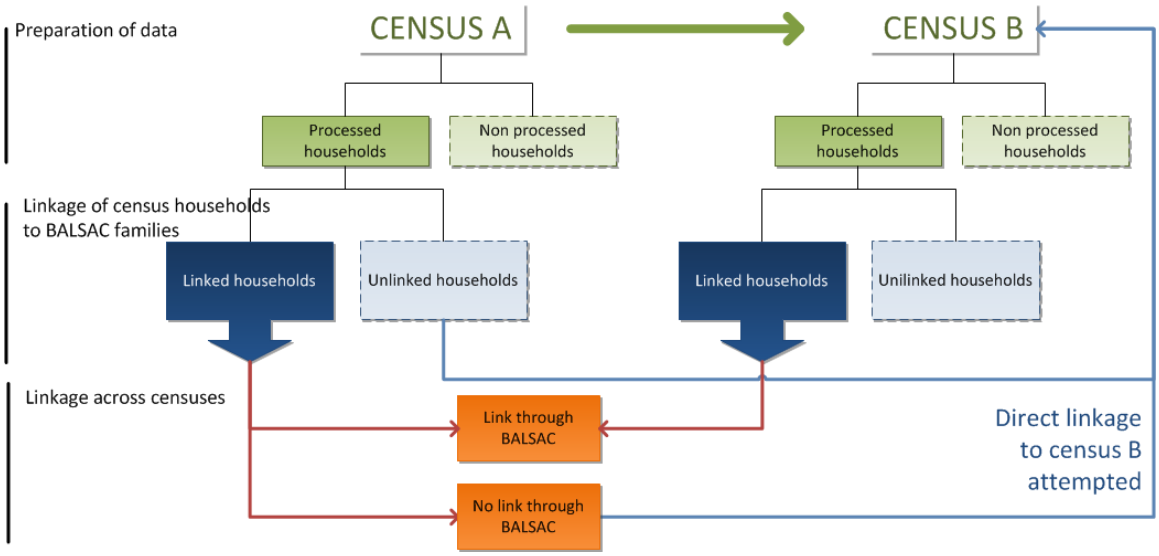

Figure 4: Linkage across censuses

An illustration of the approach: linkage results in Saguenay and Trois-Rivières

In this section, we present the results obtained on the region of Saguenay and on the city of Trois-Rivières (see Figure 1) to illustrate our linkage approach. We also carry out a critical evaluation of these results by analyzing their variability depending on characteristics of households and individuals.

The Saguenay region is situated about 200 km north of Quebec City. Settlement by French Canadians began in the second quarter of the 19th century and the population grew rapidly mostly through natural increase. At the beginning, the region consisted almost exclusively of farmers and industrialization began slowly with the implantation of pulp industries at the turn of the 20th century. Cities of modest size were in place from the late 19th century. Trois-Rivières, located on the north shore of the St. Lawrence River about halfway between Montreal and Quebec City, is a middle-sized city also populated by a majority of French Canadians. The development of hydroelectric and forest resources caused a progressive industrialization movement in the second half of the 19th century.

Census data for the two areas were digitized for the 1852-1911 Canadian censuses and Table 1 shows the number of households and individuals as well as the mean number of individuals per

household for each census. The population was slightly bigger in Saguenay already in 1852 and as the rate of increase in this region was also higher the difference in population size became more marked with each census. The mean size of household was also more important throughout the whole period in Saguenay where households contain on average one more member than in Trois-Rivières.

	Saguenay			Trois-Rivières		
	HH	**IND**	**Mean HH size**	**HH**	**IND**	**Mean HH size**
1852	681	5,031	7.4	787	4,873	6.2
1861	1,075	8,825	8.2	992	6,927	7.0
1871	1,994	11,814	5.9	1,714	8,375	4.9
1881	2,074	13,810	6.7	1,760	9,296	5.3
1891	2,300	14,776	6.4	1,700	9,119	5.4
1901	2,669	16,365	6.1	2,223	12,195	5.5
1911	3,786	23,769	6.3	2,993	16,524	5.5
Total	**14,579**	**94,390**	**6.5**	**12,169**	**67,309**	**5.5**

HH: Households; IND: Individuals

Table 1: Distribution of households and individuals in the 1852-1911 Canadian censuses for Saguenay and Trois-Rivières

As described above, linkage between census data and BALSAC was first carried out for the seven censuses. BALSAC contains all marriages celebrated in Quebec up to 1965[4] as well as birth and death records for the Saguenay region. Thus for Trois-Rivières, candidate families from BALSAC are composed of couples and their married children whereas for Saguenay they also contain all children born and/or dead in the region. In a second step, linkage to the next census was performed on households that were not already connected through BALSAC. Obviously, this type of linkage could only be done for the censuses from 1852 to 1901.

Based on our selection criteria, linkage was attempted on 97% of households in Saguenay and on 96% in Trois-Rivières. Non processed households are households with only one member, collective households (such as logging camps) and institutional households (such as religious congregations and hospitals). After performing linkage to BALSAC and to next census, we had 2.7% of processed households in Saguenay and 10.5% in Trois-Rivières that could not be linked at all. This means that overall, 94.5% of households in Saguenay and 85.8% in Trois-Rivières were linked to BALSAC or to the next census or in most instances to both sources.

Each time a household is linked either to BALSAC or to the next census, the linkage program offers the possibility to link household members individually. In the end, 91.4% of individuals in Saguenay and 61.5% in Trois-Rivières were linked to BALSAC. Among individuals in the 1852-1901 censuses, 60.4% in Saguenay and 38.1% in Trois-Rivières were linked to the next census.

[4] BALSAC contains all Catholic marriages up to 1965 and approximately 20% of non-Catholic marriages. French Canadians were Catholics and represented about 95% of the population in Trois-Rivières and 99% in Saguenay through the whole study period.

The linkage of households and individuals to BALSAC

The left panel of Figure 5 shows, for each of the seven censuses, the proportion of households that were linked to a BALSAC family. Linkage rates are above 90% (between 93 and 97%) for all censuses in Saguenay. They are lower and slightly more variable in Trois-Rivières: between 81 and 89% except for the 1871 census reaching only 75%. Differences between the two regions is certainly explained in part by the availability of birth and death records in Saguenay which facilitates the linkage (for instance in the case of very common names or surnames which would make it difficult to select a candidate family) since information will be available for a higher number of children (those who were born in the region but did not marry or married outside Quebec). Another possible explanation is that although both areas were mostly French Canadian and Catholic, there was a small non-Catholic population in Trois-Rivières (3 to 8% throughout the period) while in Saguenay the population was 99% Catholic and French Canadian during the whole study period.

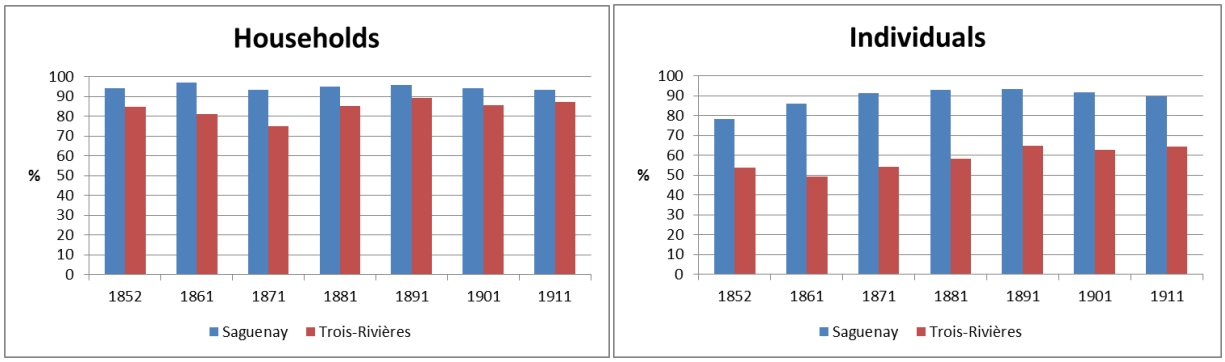

Figure 5: Proportion of households and individuals linked to BALSAC

The right panel of Figure 5 displays, for each census, the proportion of individuals that were located in BALSAC. In Saguenay, figures are slightly lower than for households (between 86 and 93%) with the exception of the 1852 where "only" 78% of individuals were linked. This could be due to the fact that as the first recorded births in Saguenay were in 1842, there were fewer individuals than for the other censuses that could be traced in BALSAC through their birth records. The linkage rates in Trois-Rivières are much lower with between 49 and 65% of individuals linked to BALSAC at each census. This is most likely explained again by the fact that when data is available on married children only, never married children or those who married outside Quebec cannot be linked.

Next, we were interested in verifying to what extent the population linked to BALSAC was different from the complete census population both at the household and individual levels[5]. In order to see what factors could impact on the probability of a household to be linked to BALSAC, we looked at linkage rates by household type, household size and occupation of head. Results are shown in Table 2. Regarding household type, substantial differences can be seen among categories in both regions with linkage rates for widows and widowers (with at least one

[5] Again we exclude here single households and collective households with a large number of members which were not considered for linkage. This represents about 3.5% of households and about 2% of individuals in all censuses.

child) 16 to 25% lower than couples. As mentioned above, the basic unit of nominative information in BALSAC are the couple's first and last names; so when we attempt linkage on households headed by a widower, we have the full name of the man but we do not have any information on his late wife's name. When we work with families headed by a widow, the situation gets more complicated because we know only the last name of the late husband and the first name of the woman which means that we miss complete names for both spouses. The section on household type in Table 2 has a category for "Other" household type which comprises mostly households where the head is said to be married but no other married individuals are found and where there is at least one child. Some of these household heads are most likely widows and widowers while others could be married individuals whose spouse was enumerated elsewhere or not at all.

	Saguenay		Trois-Rivières	
	Linked	**n**	**Linked**	**n**
Household type				
Couple	98.1%	13,101	90.7%	10,194
Widower + child	91.7%	411	79.3%	397
Widow + child	83.0%	448	69.7%	864
Other	82.8%	198	54.0%	213
Household size				
2	90.7%	1,090	77.9%	1,493
3	94.5%	1,431	83.4%	1,637
4	95.7%	1,550	87.5%	1,656
5	97.2%	1,639	89.2%	1,639
6-9	98.7%	6,075	92.3%	4,192
10-19	99.1%	2,345	93.9%	1,022
Occupation of head				
Higher occupations	90.9%	894	84.7%	2,300
Skilled workers	97.1%	1,253	89.8%	2,285
Farmers	98.9%	7,029	92.5%	949
Lower skilled workers	93.8%	513	91.2%	1,163
Unskilled workers	97.4%	3,120	92.1%	2,965
Total	**97.2%**	**14,158**	**88.1%**	**11,668**

Table 2: Proportion of households linked to BALSAC by household type and size and by occupation of head

Household size has a clear effect on linkage rates as the proportion of linked households increases with the number of people in the household. This increase is more marked among households of smaller size (2 to 5 members) but it is also apparent among larger households. This size effect is more important in Trois-Rivières (rates going from 78% in household with two members to 94% in household with 10 to 19 members) than in Saguenay (rates going from 91% in household with two members to 99% in household with 10 to 19 members). Thus the difference in linkage rates between the two region is more important in smaller households (13

percentage points in households with 2 members) than in larger ones (5 percentage points in households with 10-19 members) As the average household size is higher in Saguenay (6.5 individuals; 40% of households with 2 to 5 members) than in Trois-Rivières (5.5 individuals; 55.3% of households with 2 to 5 members), this has an effect on the difference in the global linkage rate between the two regions.

Lastly, we examined the linkage rates among occupational class of the household head[6]. As shown on the bottom part of Table 2, occupation of the household head does not seem to impact very much on the linkage rate except for the higher occupation group where we observe slightly lower rates in both areas. Factors that could explain this are either a higher proportion of non-Catholic within that group, or a higher proportion of immigrants (married outside Quebec) or a smaller mean household size.

	Saguenay		Trois-Rivières	
	Linked	**n**	**Linked**	**n**
Sex				
Male	91.4%	48,341	61.6%	31,726
Female	91.5%	44,776	61.3%	33,559
Age group				
0-9	93.3%	31,648	50.0%	17,345
10-19	89.4%	22,129	53.1%	14,907
20-29	88.2%	14,976	61.4%	11,080
30-49	93.6%	16,308	78.3%	13,409
50 and over	91.0%	8,224	74.3%	8,385
Marital status				
Single	89.3%	61,310	44.1%	40,918
Married	96.5%	29,715	88.1%	21,765
Widow-er	80.9%	2,251	60.1%	2,589
Total	**91.4%**	**93,295**	**61.5%**	**65,309**

Table 3: Proportion of individuals linked to BALSAC by sex, age group and marital status

Once a household is linked to BALSAC, the linkage of its individual members is also attempted. To assess the linkage results at this level, we have looked at linkage rates by sex, age and marital status (Table 3). Both sexes show the same linkage rate. This is very encouraging, since one might expect that female first and last names are more loosely registered than those of males, which is obviously not the case. There are however more differences between age groups. They are less marked in the Saguenay region, thanks again to the availability of birth and death certificates, but they are sizable in Trois-Rivières. There, linkage rates increase with

[6] Occupations from the census were coded in HISCO (van Leeuwen and Maas 2002) and classified according to HISCLASS (van Leeuwen and Maas 2011). The five occupational classes presented here are a grouping of the 12 HISCLASS categories.

age and this is directly related to the fact that BALSAC families contain only married children. Other things being equal, the older a person, the more chances he or she has to have married and thus be present in BALSAC. After 50, the rate decreases a little due to widowed individuals. In the Saguenay region, the minor variations in linkage rates are most likely explained by migration. We see a slight decrease in the rates among individuals aged 10-29 which reveals the higher mobility of young adults who as a result have more chances to marry outside Quebec. If they were born outside the region (and thus without a birth certificate in the regional vital records), we would then miss them.

Lastly, looking at linkage rates by marital status, we can again apprehend the direct effect of the availability of birth and death records in Saguenay: in that region, the highest rate is found among married individuals followed by single individuals (who are mostly children and young adults still living with their parents since we do not link single households or households where nobody is married) and then by widowed individuals. However, the variation in rates is much less important than in Trois-Rivières where the rate is twice as high among married people than among singles.

The linkage of households and individuals to the next census (1852-1901)

Linking to the next census poses some conditions. First of all, the household or the individual must be enumerated within the same area being compared from one census to the other. Local or regional socioeconomic context will thus play a significant role in the probability to find someone on the same territory over a decade. Second, despite its plasticity -some members being gone or added during the 10-year gap between the two enumerations-, a household must present a minimal consistency to be linked (at least one common conjugal unit must be found in both censuses). Third, it has to be enumerated in such a way that the nominative information can be equally legible in both censuses. This involves a good knowledge of all household members (names, ages, marital status, and so on) by informants and an appropriate transcription of the information by census takers (continuity of personal information: last and first names spelling, ages, place of birth). If the information is too different, we might miss a true linkage because the household in the second census will not be selected by the program as potential candidate for linkage.

Notwithstanding these difficulties, linkage to the next census was attempted on all processed households in the 1852-1901 censuses that were not connected to the next census through linkage with BALSAC. The results presented in the left panel of Figure 6 include all households linked to a household in the next census either through BALSAC or through direct linkage. In Saguenay, more than 99.5% of linkages were done through BALSAC indicating that a tiny proportion of households appearing in two successive censuses did not have a head married in a Catholic parish of Quebec. In Trois-Rivières, this proportion is also very low, between 1 and 5% depending on the census. Again this could be explained by the small but not negligible proportion of non-Catholics in this city. As shown on Figure 6, results are lower in Trois-Rivières (between 49 and 66% of linked households) than in Saguenay (between 67 and 83%) but interestingly both series display a similar distribution with rates decreasing and reaching the lowest values in 1871 et 1881 and then increasing again.

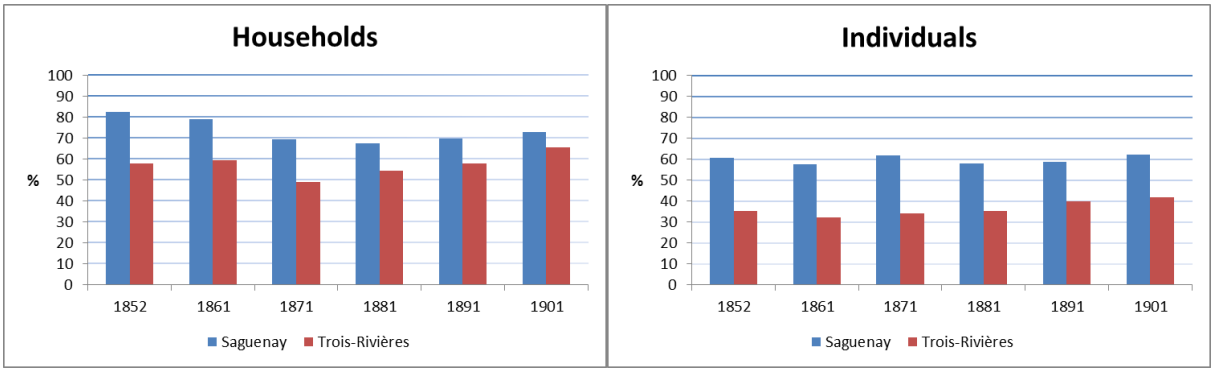

Figure 6: Proportion of households and individuals linked to the next census

As we observed in the linkage to BALSAC, the linkage rates to the next census are lower among individuals than among households in both regions. They are also more uniform across censuses (from 57 to 62% in Saguenay and from 32 to 42% in Trois-Rivières). Lower rates can of course be explained by death or migration of individuals between censuses. In the case of migration, the effect could also explain in part the differences between the two areas because Saguenay represents a larger territorial unit than Trois-Rivières (region vs single town) and thus we are mostly likely missing short distance moves in the latter. Differences in the composition of the population such as sex ratio, age distribution, size and types of households could also contribute to the observed differences. And of course, the possibility of relying on birth and death records in Saguenay would also have an effect.

Like previously done for the linkage to BALSAC, we looked at linkage rates by household type, household size and occupation of head in order to see what factors could impact on the probability of a household to be linked to the next census. Results are shown in Table 4 and present the proportions of households in Census A (see Figure 6) that were linked to Census B.

More importantly than for the linkage to BALSAC but probably for the same reasons, the household type heavily influences the probability to get a link to the next census. Compared to couples, the linkage rate for widowers decreases by 17% in Saguenay and 24% in Trois-Rivières, and by 30 and 40% respectively for the widows. These differences are also partly explained by mortality among widowers and widows who as a group are older than married individuals.

Household size plays a major role when attempting linkage of a household in two consecutive censuses. It is for a good share due to the fact that the loss of household members (by migration, marriage or death) greatly reduces the chances to find the household ten years after. Table 4 shows very clearly that the decrease in success rates is important especially for households counting two or three members. Among larger households, it is likely that migration explains most of the observed variation in linkage rates.

Lastly, linkage rates also vary according to the head's occupation[7]. Farmers, known to be more sedentary, show the highest score. Other groups in Saguenay have somewhat lower and very

[7] Note that unclassified occupations have been excluded from the calculation, explaining why the global value is slightly different from the mean of the categories.

similar linkage rates. In Trois-Rivières, lower skilled workers have slightly lower rates than farmers while the three other groups have the lowest rates. The difference between the two areas in the rates for skilled workers may be due to the progressive industrialisation process in city of Trois-Rivières which could have prompted the migration of a good share of the craftsmen. However, as a whole, the socio-occupational category displays a modest effect on success rates.

	Saguenay		Trois-Rivières	
	Linked	**n**	**Linked**	**n**
Household type				
Couple	75.6%	9,635	63.3%	7,597
Widower + child	62.7%	327	47.7%	310
Widow + child	53.4%	335	37.8%	686
Other	62.9%	151	50.6%	156
Household size				
2	58.0%	793	40.1%	1,111
3	66.3%	1,041	48.7%	1,188
4	69.9%	1,081	60.1%	1,232
5	71.2%	1,168	62.4%	1,222
6-9	78.0%	4,527	68.7%	3,185
10-19	81.2%	1,813	72.3%	787
Occupation of head				
Higher occupation	67.6%	547	62.1%	1,746
Skilled workers	71.4%	839	60.8%	1,777
Farmers	78.5%	5,858	70.0%	766
Lower skilled workers	70.1%	278	66.5%	813
Unskilled workers	68.4%	2,008	61,2%	2,179
Total	**74.3%**	**10,448**	**60.5%**	**8,749**

Table 4: Proportion of households linked to the next census by household type and size and by occupation of head

At the individual level (Table 5), the overall linkage rate between censuses corresponds to approximately two thirds of the linkage rate to BALSAC (66.1% in the Saguenay, 62.0% in Trois-Rivières). As noted for the linkage to BALSAC, the rates are very similar for men and women while age and marital status are important factors of differentiation.

Similar to what we observed in the linkage to BALSAC, the 30-39 age group has the highest linkage results with the difference between Saguenay and Trois-Rivières probably explained by the size of the enumerated territory. The linkage rate among people aged 50 and over is much lower however than their linkage rate to BALSAC and this is evidently due to mortality. The decrease in the linkage rate for the 0-9 age group is smaller in Trois-Rivières than in Saguenay (25% compared to 37%). Since we link many children through their birth in Saguenay, we have a higher probability to lose them because of death or migration before the next census than in Trois-Rivières where we can only link children who have survived to adulthood and married in

Quebec. For the two next age groups, the losses are similar in both areas (34 to 38%) and likely reflect the effect of migrations, a part of them taking place outside the province. Finally, the general pattern for the age groups is also perceptible when considering the marital status: a better linkage rate for the married (especially in Trois-Rivières), and a strong effect of mortality for widowed individuals.

	Saguenay		Trois-Rivières	
	Linked	**n**	**Linked**	**n**
Sex				
Male	61.3%	36,320	38.7%	23,922
Female	59.5%	33,454	37.6%	25,387
Age group				
0-9	64.6%	23,643	37.2%	13,229
10-19	56.3%	16,935	32.4%	11,498
20-29	58.6%	11,257	37.6%	8,218
30-49	66.1%	12,109	50.0%	10,020
50 and over	47.5%	5,973	33.0%	6,197
Marital status				
Single	58.7%	46,317	32.0%	31,263
Married	66.0%	21,945	52.2%	16,149
Widow-er	34.2%	1,649	19.7%	1,903
Total	**60.4%**	**69,922**	**38.1%**	**49,332**

Table 5: Proportion of individuals linked to the next census by sex, age group and marital status

A special case: the linkage of women across censuses

Figure 7 shows the distribution of the 31,351 females and 31,389 males who were linked to BALSAC according to the number of censuses where they appear. About 48% of individuals in Saguenay and 52% in Trois-Rivières appear in only one census. Among them, half are individuals who appear for the first time in the 1911 census and thus could not have been linked to another census. Approximately one quarter of individuals linked to BALSAC appear in two censuses. In Saguenay, another 25% appears in three to five censuses whereas this figure is about 21% in Trois-Rivières. Lastly, 863 individuals in Saguenay and 258 in Trois-Rivières appear in 6 censuses and 208 in Saguenay and 72 in Trois-Rivières were found in the 7 censuses. The proportion of individuals linked to more than two censuses is slightly higher in Saguenay in line with the higher global rate of linkage.

But the most remarkable feature of these results is the similarity of results obtained on men and women in both regions. In Saguenay, among individuals linked to more than one census, we were able to find 2,595 men and 3,141 women whose marital status changes from single to married. This is also the case for 916 men and 1,056 women in Trois-Rivières. Thanks to the connection with BALSAC, women, just like men, can be located in their family of origin and in their conjugal family allowing us to follow them throughout their life course in all censuses

where they appear. The fact that we can trace both women and men points to the fact that our linkage approach overcomes some difficulties related to the nature of data and shows how complementary the two data sources are.

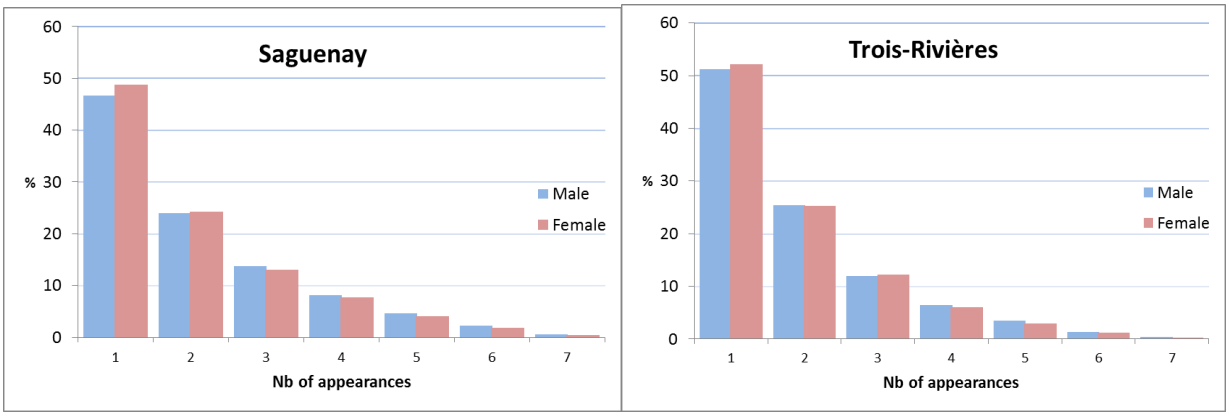

Figure 7: Distribution of census individuals who were linked to BALSAC according to the number of censuses where they appear

Discussion

Possibilities and limits of the approach used for linkage

Our linkage approach is based on the comparison of two sets of nuclear families: the first one reconstituted from vital events in the BALSAC database and the second one extracted from households in Canadian historical censuses. It relies mostly on nominative information but contextual information on family members' characteristics is also used. Our objective is to link as many families and individuals from the census population as possible with the most efficient method in terms of reliability and efficiency using a program developed by our group.

Because of the differences in the two data sources, the program does not allow for a perfect match, but rather a match with various degrees of similarity between BALSAC families and census households. Family structures from BALSAC may be incomplete (in fact except for Saguenay they contain only couples and their married children). Census households can contain various types of families (nuclear with one or two parents, recomposed, multigenerational) and include members outside the family. The program can deal with situations where the nuclear family cannot be fully restored as long as the structure is not completely fragmented. The success rate is obviously dependent on the degree of alteration of the structure.

From the viewpoint of census data, the linkage with civil records, especially marriages, is of tremendous help to link censuses and build census-based longitudinal datasets. Through the BALSAC family file, it is possible to connect a single male child within his parents' household and the married man appearing as a household head ten years later. Marriage records do not only confirm links based on individual's nominative information, but are also necessary to solve a large number of homonymy problems. In addition, due to the loss of maiden names after marriage, they are the only source which could allow the linkage of a girl within her parents'

household with the woman being a member of a couple in a subsequent census. Results on Saguenay and Trois-Rivières show that our linkage method allows for reconstruction of the lifecourse of both men and women from childhood to adulthood which is quite unique in a North American context.

Considering couple (or couple and child) mentions as the unit of comparison, preferably to a unit based on individual information, minimizes the negative impact of nominative variations on linkage performance and increases success rates. Using this approach, linkage will be attempted on all households with at least two members except those for which it is impossible to compose at least two nominative mentions, and those whose members are all single (such as religious communities, hospitals, etc.). Thus, our linkage procedures will yield longitudinal data biased in favour of couples and families. However we have seen that unprocessed households in Saguenay and Trois-Rivières represent less than 5% in the seven censuses 1852-1911.

We use the BALSAC population database as the reference data for vital events. The advantage of using BALSAC is that it covers the whole territory of Quebec for the entire period for which nominal censuses are available and it is fully family-structured. Except for the Saguenay region, it contains at the moment only marriages so families are composed of parents and their married children in most instances. This also plays in favor of more linkage for couples and families. At the individual level, married individuals will have an advantage over single and widowed persons. The differences between the three types of marital status are more marked for Trois-Rivières showing the impact of having access to birth and death records especially for linking children.

Some possible biases could also be brought about by our linkage method. There as many nominative units of comparison in a family as there are children therefore the probability of linkage increases with the number of children. Results show that household size has a strong effect with the probability of linkage increasing with the number of household members. Moreover, we do not rely on places of residence for the selection of candidate families but we use it in the calculation of the score to rank the candidates. Thus stable families might have a certain advantage over those who move.

When a household is linked, the program enables the linkage of individuals. However, the correspondence of individuals is more difficult to establish and following individual fates is complex. From a census to the next, for example, an individual child can either remain in the parental household, or be part of another type of household as a single-adult or married-adult (in the case of girls, they will then lose their name to take the name of their husband), or be lost to observation because of death or migration. Moreover, because the linkage methodology focuses on the selection and extraction of nuclear families, household members who are not part of this type of family unit are less easily included in the process.

Notwithstanding these limits and potential biases, we have obtained very high linkage rates to BALSAC especially in Saguenay and to a lesser extent in Trois-Rivières. Of course, since in Saguenay we rely on the complete set of vital events, we expected these higher success rates and they clearly demonstrate that availability of birth and death records allows the program to produce optimal results. But we have seen that other factors can explain the differences

between the two areas such as the size of the territory and its composition (modest size cities and rural areas in Saguenay vs mostly urban setting in Trois-Rivières) or some characteristics of the population like a mean household size higher in Saguenay or the presence of a small non-Catholic population in Trois-Rivières. Occupational structure also plays a role since farmers have higher rates of linkage and are far more numerous in Saguenay than in Trois-Rivières while the higher occupation group has lower rates and is more represented in Trois-Rivières.

Data available for research in the IMPQ

Once completed, the infrastructure will make the following data available for research: 1) all marriage certificates since the implementation of parish registers in Québec in the early 17th century to 1965. This will enable automatic reconstruction of the genealogy of the Quebec population over a period of three-and-a-half centuries; 2) linked data based on three types of vital event records (birth, marriage and death certificates) from the beginning of the 17th century to 1849. The 2.3 million certificates will allow researchers to explore the historical demography of families over a 250-year period; 3) complete-count census microdata covering two urban settings (Quebec City and Trois-Rivières) and three regions mixing rural and urban environments (Gaspésie, Côte-Nord and Saguenay) across seven modern nominal censuses (1852 to 1911). These microdata (close to a million individuals belonging to 161,000 distinct households) will be linked across censuses and to corresponding vital event data.

This vast array of biographical information will permit studies based on individual trajectories situated within families, households and communities and examined from a multigenerational perspective. In the North-American context, such an infrastructure is going to be unpaired at this scope[8]. Based upon nominative data, it will easily support any linkage operation of other serial sources, such as assessment rolls, city directories, or notarial archives, to conduct in-depth studies in various fields of social sciences and population genetics.

Emerging research opportunities

The interconnection between civil records and the censuses, along with linkages across the censuses themselves, will substantially broaden and enrich the avenues of research in both the social and the biological sciences. It will be possible to conduct detailed studies on a crucial period in the history of the Quebec population (mid-19th century to the first decade of the 20th century) focussing on the evolution and long-term consequences of phenomena such as cultural diversification, social mobility and intercommunity relationships. During this period Quebec's inhabited space[9] expanded as a result of new agricultural settlements and broadened maritime and forestry activity; at the same time, Québec was transitioning to an industrial economy, and urbanization was accelerating. One of the most spectacular results of the linkage process will be the capacity to update the too often overshadowed life histories of one half of the population: women.

[8] In Canada, the only comparable dataset (but not publicly accessible) is the sample of Montreal dwellers gathered by Sherry Olson and Patricia Thornton for the period 1861-1901 and largely supporting their analyses in Olson and Thornton, 2011.

[9] We mean the territory occupied by sedentary populations of mostly European descent.

From a population genetics and biodemographical perspective, the infrastructure will considerably enrich research on the transmission of biological and sociocultural characteristics, on the genetic diversity in Quebec regional populations and on the factors that have shaped this diversity. These will translate into original studies that can contribute to a better understanding of the genetic determinants of health.

Concluding remarks

Our goal was to develop tools and procedures for systematic and automated linkage involving vital records and census data. Development and testing operations are largely completed and we now have at our disposal a powerful program providing the ability to select the desired type of linkage (census to BALSAC or census to census). Moreover once a household is linked it is possible to link its members.

Our linkage results are variable depending on the region and on the census year and further work could certainly be done to attempt manual linkage on individuals and households that could not be linked with the program. However, already we can provide to the research community original datasets on the Quebec population combining information on the two types of data which is unprecedented. In the course of the IMPQ project, a portal will be developed to provide access to these data.

Linkage with BALSAC and across censuses is ongoing on three more regions where we have complete count census data. We also pursue our work on the description and analysis of the linkage results .A next step will be to perform logistic regressions to specify the role of different factors on the probability of linkage. We suspect that some of these factors -such as household size and occupation- are related and we want to analyse the net effect of each of these variables. Moreover, we intend to look more closely at unlinked households to get a better understanding of their characteristics.

Lastly, the software could also be improved. It would be desirable to come to greater automation of decision-making and to further refine the process of individual linkage. But even it its current state it provides an innovative tool with the ability to combine longitudinal information from BALSAC to census data and to create census-based longitudinal datasets.

Acknowledgements

This project is funded by the Canadian Foundation for innovation, the Quebec ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie as well as the Université du Québec à Chicoutimi, Université du Québec à Trois Rivières and Université de Montréal.

Other researchers members of the IMPQ project include Danielle Gauvreau (Concordia University), Bertrand Desjardins, Lisa Dillon, Alain Gagnon, Damian Labuda (Université de Montréal), Richard Marcoux (Université Laval), France Normand (Université du Québec à Trois Rivières), Marc Trembaly (Université du Québec à Chicoutimi). We also acknowledge the support of research professionals, clerks and technical staff especially Mario Bourque, Dominic Desbiens and Ève-Marie Lavoie.

Bibliography

Bellavance C, Normand F, Ruppert ES (2007) Census in Context: Documenting and Understanding the Making of Early-Twentieth-Century Canadian Censuses. *Historical Methods,* 40, 2: 92-103.

Bouchard G, Bouchard M (1981) *Ajouts au code de transcription phonétique FONEM*. Document de travail BALSAC II-C-66.

Bouchard G, Brard P, Lavoie Y (1981a) FONEM: un code de transcription phonétique pour la reconstitution automatique des familles saguenéennes, *Population*, no. 6 (novembre-décembre), pp. 1085-1104.

Bouchard G, Casgrain B, Roy R (1981b) *Tests de validation des fiches de couple par ordinateur*. Document de travail BALSAC II-C-67.

Bouchard G, Larose A (1976) La réglementation du contenu des actes de baptême, mariage, sépulture, au Québec, des origines à nos jours. *Revue d'histoire de l'Amérique française*, 30-1: 67-84.

Bouchard G, Roy R, Casgrain B (1985) *Reconstitution automatique des familles. Le système SOREP.* 2 volumes. Chicoutimi, Université du Québec à Chicoutimi, 745 pages.

Charbonneau H, Lavoie Y, Légaré J (1970) Recensements et registres paroissiaux du Canada durant la période 1665-68. Étude critique. *Population*, 25, 1: 97-124.

Gauvreau, D (1991) *Québec, une ville et sa population au temps de la Nouvelle-France*. Sillery, Les Presses de l'Université du Québec.

Gauvreau D, Thornton P, Vézina H (2010) Le jumelage des recensements aux mariages du fichier BALSAC présentation de l'approche et étude exploratoire des enfants de couples mixtes à la fin du XIXe siècle. *Cahiers québécois de démographie,* 39: 357-381.

Gossage P (1999) *Families in Transition: Industry and Population in Nineteenth-Century Saint-Hyacinthe.* Montreal/Kingston, McGill-Queen's University Press.

Olson S, Thornton P (2011) *Peopling the North American City: Montreal 1840-1900.* Montreal, McGill-Queens Press.

St-Hilaire M (1990) *Description du format d'impression d'une fiche de couple*. Document de travail BALSAC I-C-96.

St-Hilaire M (2009) Filling the gap: The use of marriage records to help with inter-censuses linkage of young adults in Quebec City (1851-1911). Paper presented at the 2009 *RECORDLINK* workshop, University of Guelph.

St-Hilaire M, Marcoux R (2006) Dynamique démographique dans une capitale en reconversion économique : Québec, 1851-1901. Communication au congrès de la Société historique du Canada (London, Ontario).

St-Hilaire M, Richard L, Marcoux R (2014) Individual and familial lifecourses in Quebec City, 1871-1911: Some considerations on two biographical data-sets. In G Darroch (ed.), *The Dawn of "Canada's Century:" The Hidden Histories*, Montréal-Kingston, McGill-Queen's University Press, 322-358

St-Hilaire M, Vézina H (2010) Between household and family: The use of marriage records to link census data (Quebec City, 1852-1911). Paper presented at the 2010 *RECORDLINK* workshop, University of Guelph.

van Leeuwen MH, Maas I (2011) HISCLASS A Historical International Social Class Scheme

van Leeuwen MH, Maas I , Miles A (2002) *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven University Press

Vézina H, St-Hilaire M (2011) Entre famille et ménage : le jumelage des données d'état civil et de recensement dans la population québécoise. Colloque « La fin des recensements ». 79e congrès de l'ACFAS, Sherbrooke, 9-13 mai.