Following individuals across seven censuses: a validation tool for linkage results

Jean-Sébastien Bournival
Projet BALSAC, Université du Québec à Chicoutimi

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

Hélène Vézina
Projet BALSAC, Université du Québec à Chicoutimi

Paper presented in the session *Innovative techniques for record linking* at 41th Annual meeting of the Social Science History Association, 17 November 2016.

Abstract

In previous work, we have exposed the method and program used to match microdata from the Canadian censuses to Quebec vital records. Linkage operations rely heavily on nominative information known to be reliable in both sources and accordingly, high linkage rates were obtained. However, the quality of the data does not guarantee that the links are good and validation and consistency checks are essential. In this presentation, we show how the final step of our linkage strategy which involves the combination of linkage results on individuals and families across multiple datasets can serve as a powerful validation tool.

By matching the two sources together and by linking across censuses, the linkage process generates multiple links on a single individual since at each census an individual can be linked to BALSAC and to the next census. This set of independent links is then pooled and used to reconstruct individual biographies based on the sequence of appearances in the censuses. By combining the links, we can increase the number of links since for instance a match between two censuses can be missed but be found through linkage to the same individual in BALSAC. Even more importantly, the analysis of these sequences also allows for detection of inconsistencies and duplicates and can thus be used as a tool for validation and correction of erroneous links. We will provide an illustration of specific cases observed in the Saguenay population for the 1852-1911 censuses.

Introduction

In previous work, we have exposed the method and program developed at BALSAC[1] to match microdata from the Canadian censuses to Quebec vital records in the context of the ongoing construction of the Integrated Infrastructure of the Quebec Population Historical Microdata (IMPQ) (Vézina, St-Hilaire, & Bournival, 2015). Here, we report on the final step of our linkage strategy which involves the combination of linkage results on individuals and families across multiple datasets and we show how this last step also serves as a powerful validation tool.

As shown in a previous study, our linkage operations rely heavily on nominative information known to be reliable in both sources, and accordingly, high linkage rates were obtained (Bournival, St-Hilaire, & Vézina, 2016; Vézina et al., 2015). However, the quality of the data does not guarantee the accuracy of the links created and consequently validation and consistency checks are essential to detect ambiguous links, make corrections and evaluate the error rate.

Depending on the linkage method used, the validation process can be achieved in different ways. For automated approaches often based on a probabilistic linkage procedure, the aim is to reduce to a minimum the creation of false associations (false positives) and the omission of true links (false negatives). This kind of approach has the advantage of clearly stating the error rate generated by the linkage process (e.g.: Antonie, Inwood, Lizotte, & Andrew Ross, 2013; Goeken, Huynh, Lynch, & Vick, 2011). In a sense, the validation step is embedded in the process itself. However, with manual or semi-automated approaches, such as ours, that rely at least in part on human intervention and decision-making, the validation process implies the verification and correction of created links. The validation step is therefore performed after the linkage operations have been carried out.

The type of data used has also a great impact on the validation process. The linkage between censuses does not offer the same level of confidence than linkage between censuses and parish registers. In the first case, context relies only on census information and in order to avoid selection biases, the number of variables taken into account can be very limited (Ruggles, 2002; Wisselgren, Edvinsson, Berggren, & Larsson, 2014). In the second case, various variables from both sources can be used either to be included into linkage algorithms or simply to compare both entities (family vs household) during the matching process. More specifically, the comparison of complete family files such as those found in the BALSAC database and nuclear-type households can greatly help the linkage process.

[1] For more information on BALSAC see http://balsac.uqac.ca/english/.

First, it is necessary to briefly recall the purpose of the linkage in the context of the IMPQ creation and how the linkage program works. The main goal is to connect as many families and individuals as possible between the seven Canadian censuses spanning 1852-1911 and the civil records found in BALSAC to achieve the greatest level of integration of the two sources. We consciously favored completeness over representativeness in order to produce deeper and richer biographies. Here, it must be kept in mind that the vital events found in BALSAC are already fully linked giving access to individual biographies and family histories. This makes a huge difference in the approach since we can benefit from the rich family context to link individuals in the censuses. Moreover, family reconstitutions in BALSAC were completely validated and several consistency checks were performed to ensure that individuals and families had coherent histories.

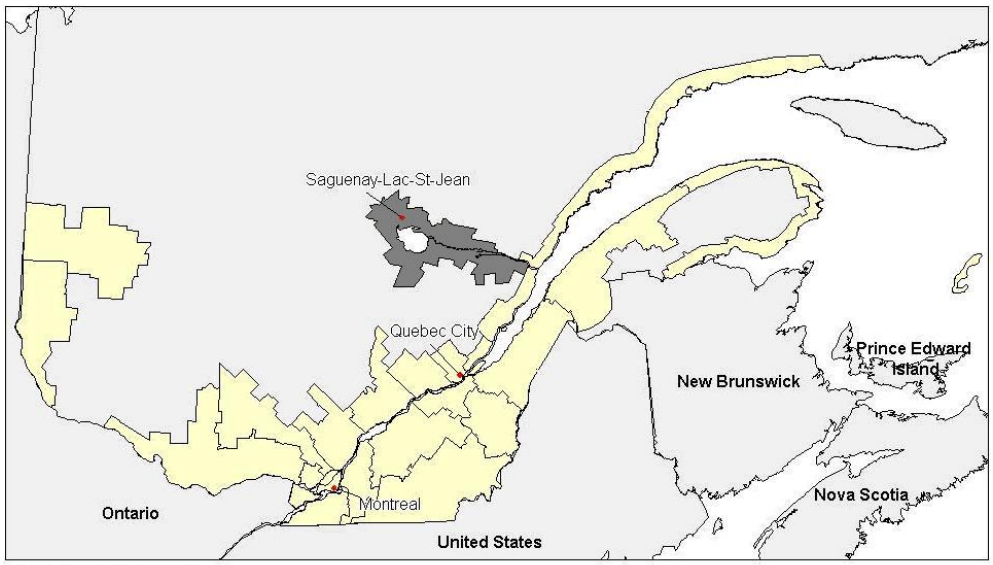
Our linkage operations rely on two independent modules. The first one is used for linking census data to civil records while the second one allows for linkage across censuses. Each of these two modules relies on the same basic principle: the comparison of the nominative elements of census household heads to couples in BALSAC or to household heads in the following census. Comparison and decision algorithms are used to select candidates and to generate scores that classify the potential candidates. In order to be considered for linkage, a household must contain at least two members forming a nuclear-type familial unit (husband and wife or widow-er with a child) since a minimum of three nominative elements is needed. Therefore, lone-headed households, institutions, working places such as lumber camps as well as parentless sibships are excluded from the process since they do not fit the selection criterion. These households, as well as their members, are classified as unprocessed.

By matching the two sources together and by linking across censuses, the linkage process generates multiple links for a single individual since at each census an individual can be linked to BALSAC and to the next census. This set of independent links is then pooled and used to reconstruct individual biographies based on the sequence of appearances in the censuses. By combining the links, we can further increase the number of links since, for instance, a match between two censuses can be missed but be found through linkage to the same individual in BALSAC. Even more importantly, the analysis of these sequences allows for detection of inconsistencies and duplicates and can thus be used as a tool for validation and correction of erroneous links. In the course of the IMPQ construction, we have now completed the linkage work on the Saguenay population for seven censuses (1852-1911). Here we will describe our approach for pooling and validating linkage results on individuals and provide an illustration of the process using the Saguenay data.

Data

The Saguenay-Lac-St-Jean region region is located 200 km north of Quebec City (Fig. 1). French Canadian settlers arrived in the 1830s, but the first vital event officially recorded was a marriage celebrated in 1842. At the beginning, the area consisted almost exclusively of farmers who came mostly from the nearby area of Charlevoix (Bouchard & Larouche, 1988). By 1852 the regional population was around 6,000 individuals and reached 37,000 50 years later. Most of the population remained rural up to the 1930s. Until the end of the 19th century, the majority of farmers alternately worked on the family farm during the summer and logged the forests during the winter to maintain incomes flow. In 1901, Chicoutimi, the main regional center, barely exceeded 5,000 inhabitants but managed to evolve as an urban pole as it benefited from diverse regional functions (administration, hospital, courthouse, etc.). It is not before the Second World War that the population became predominantly urban.

Figure 1. Location of the Saguenay-Lac-St-Jean region (Quebec, Canada)



Source: GRIG, UQAC and CIEQ-Laval

As mentioned above, our project aims at comparing two sources of data. First, we rely on vital events from the BALSAC database maintained at the Université du Québec à Chicoutimi. Created more than 40 years ago, the database contains all the Catholic marriages since the early days of French settlement in the 17th century to 1965 (1971 for Saguenay). Births and deaths records are currently available only for the Saguenay region but ongoing work aiming at the creation of the IMPQ will make available births and deaths records for the whole province of Quebec up until 1850. Currently, BALSAC contains 2.9M records which were linked according to the family reconstitution method

inspired by Louis Henry (Fleury & Henry, 1956), thus providing information on 2.5M families and 5M individuals.

The second set of data comes from the seven Canadian historical censuses between 1852 and 1911. We use full count data (100%) for the Chicoutimi census district which corresponds to the part of the region located along the Saguenay River. Household and individual counts per census are found in Table 1. Altogether, our dataset comprises almost 95,000 individuals distributed in 14,579 households.

Table 1. Distribution of households and individuals found in the Chicoutimi district in the 1852-1911 censuses and proportion (%) linked to BALSAC and to the next census

Census year	Households	% linked to BALSAC	% linked to next census	Individuals	% linked to BALSAC	% linked to next census
1852	681	93.5	81.9	5031	78.1	63.7
1861	1075	96.8	79.2	8825	85.8	60.3
1871	1994	93.5	69.4	11,814	91.2	64.6
1881	2074	95.0	67.9	13,810	92.7	59.9
1891	2300	95.9	70.2	14,776	95.1	60.9
1901	2669	93.9	72.8	16,365	91.5	64.0
1911	3786	93.3	-	23,769	89.7	-
Total	**14,579**			**94,390**		

Table 1 also displays for each census the linkage rates to BALSAC and to the next census. Obviously, being the last one, 1911 cannot be linked to another census. For households, linkage rates to BALSAC are high, pointing to the complementarity of the two sources and to the quality of the data (Bournival et al., 2016). Unlinked households are mainly those which do not have the nuclear-type structure necessary to fit the linkage program, mostly institutions, small shops and solitary households[2]. The linkage rate is slightly lower and more variable for individuals, especially for the 1852 and the 1861 censuses. This situation is largely explained by the lower number of birth and death records for individuals found in these censuses (since settlement started in the 1830's). We have shown previously that access to births and deaths records makes a significant difference in the linkage successes (Vézina et al., 2015).

Rates in census to census linkage are lower both at the household and the individual levels. Among individuals, rates are rather uniform across censuses. For households, rates are first decreasing reaching the lowest values in 1871 and 1881 and then increasing again. Households must present a minimal consistency to be linked (at least one common

[2] For more information on the household selection for the linkage process see (Vézina et al., 2015).

conjugal unit must be found in both censuses) and this consistency depends on the incidence of deaths and migrations between censuses at the individual level.

The validation process

Once the two datasets are linked together and beyond the success demonstrated by the high matching rates, a question remains: "How do we know that these links are correct?" A validation process is necessary to highlight and to correct potential mismatches that occurred during the linkage process as well as to calculate error rates.

One way to validate the linkage results could be to sample a fraction of all the links (e.g.: one-fifth or one-tenth), to attempt linkage a second time and to compare the two sets of results. However, as straightforward as it can be, a typical validation method like this is not well suited for the linkage process used at BALSAC. We use a computer-assisted linkage method that generates a pool of candidates using information contained in the censuses (names of the couple to match, names of the children, place of residence). A score is attributed to each candidate couple found in the BALSAC database. Couples are then ordered by score and only those with the highest scores are retained. In Saguenay, the linkage between census households and BALSAC has shown that in 75% of cases the candidate selected for linkage was the one with the highest score. Moreover, in many instances, the highest score is significantly higher than the second one, strengthening the confidence in the choice proposed by the program. If we were to attempt linkage a second time on a subset of data, the pool of candidates would be exactly the same leading most of the time to the same linkage decision as no new information could be introduced to challenge the first decision. Therefore, a validation step at this stage of the process would be useless.

Before we examine another way to validate the linkage results, it is worth mentioning that a number of characteristics of the linkage process, without being validation tools as such, greatly contribute to reinforce confidence in the created links and to decrease the risk of erroneous links. First, working with the civil records linked in BALSAC facilitates the matches. As noted earlier, we use family and individual biographies to locate people in space (with the help of location variables available from both sources) and in time (thanks to dates and age). Furthermore, for each individual, BALSAC contains all nominative variations observed across vital events of which he or she was part[3]. It is not uncommon to find name variations across events sometimes because of a registration error (misspelling), sometimes because a person uses two or more names

[3] The treatment of names was developed for the BALSAC database in the early 1980s. The linkage between censuses and civil records relies on all the work done to create dictionaries of names that account for inversions, equivalences and orthographic variations. For more information on this subject see (Bouchard, Brard, & Lavoie, 1981; Bouchard, Roy, & Casgrain, 1985; Bouchard & Pouyez, 1980).

interchangeably. It is then possible to link people with different names (nicknames or equivalences) in successive censuses. We must emphasize that the linking process relies on the use of phonetic representations as a way to standardize names, a process that has proved helpful in matching individuals (Vick & Huynh, 2011). This is true at least for individuals enumerated in their family household. Individuals enumerated outside their family household remain harder to match due to the lack of contextual data and to a high level of homonymy in Saguenay. For example, 2% of men and women shared the same name in the 1891 Saguenay census. There were around 140 “Joseph Tremblay” and “Marie Tremblay”, a number of which could not be linked despite all the efforts taken to maximize possible matches (Bournival et al., 2016). Another characteristic of our linkage program which increases our confidence in the links created is that matching rests on the comparison of couples instead of individuals which greatly helps to reduce ambiguous cases. A last one is that family heads, their children and other members of the household can be linked simultaneously. The math is simple: the more nominative information is added in the comparison process the more discriminant it becomes.

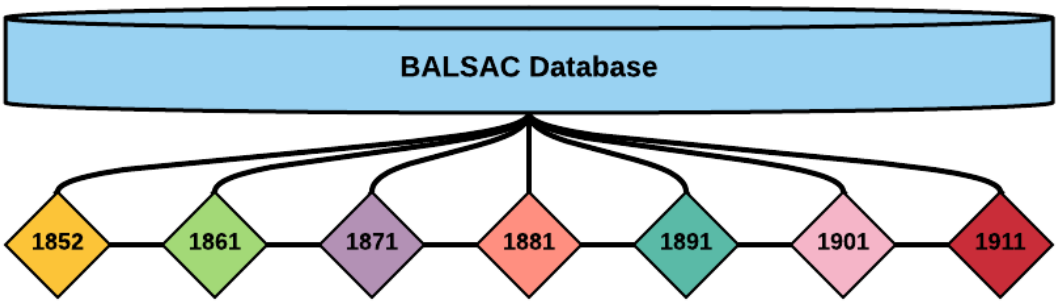
Thus due to the high quality of the datasets and their level of complementarity as well as to the characteristics of the linkage process described above, we can confidently follow Wisselgren et al. (2014, p. 141) and consider that we create “secure links” in the vast majority of instances. However, our method does not completely prevent problems such as homonymy and errors can occur through the process. Migration, death, as well as variations in household composition over time also sometimes create ambiguities and prevent linkage.

In the next section, we will show that the links created by pooling the results obtained during linkage operations can represent a useful tool to detect inconsistencies in individual biographies. Moreover, we will see that most ambiguous cases can be recovered if they were missed in the first place or corrected if erroneous.

Combining the linkage results to create individual sequences

Each of the two linkage module generates its own type of links: the first module links census data to BALSAC civil records, the second links data from two distinct censuses. When performing linkage on a given population, we start with the first module and proceed with the seven censuses. Once this is done, we rely on the second module to link censuses pairwise and links are created independently of matches made with the first module, thus ensuring unbiased decision-making. Each census is linked forward to the next one in chronological order so that when the work is completed we have six pairs of censuses linked together (see Figure 2 below).

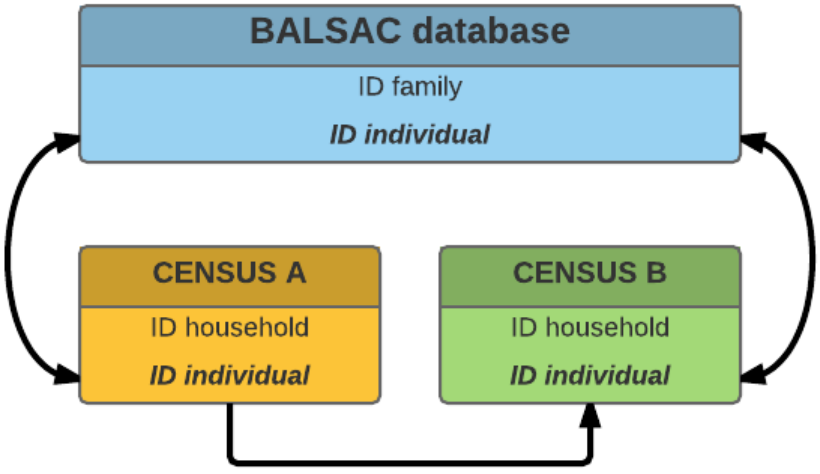
Figure 2. The 13 linkage operations performed between BALSAC and the 7 Canadian censuses



As the linkage program relies on nuclear-type family structures, families must first be extracted from households in order to match them with families in the civil records (Vézina, St-Hilaire, & Bellavance, 2014). Each census individual linked to the BALSAC database is connected to an individual ID number and to the ID number of the family where he is found at the time of the census (his parents' or his own if he is married). Each BALSAC ID number is unique meaning that individuals and families in different censuses who were linked to the same individual or family in BALSAC can be traced and connected across censuses. This also implies that linkage to the same individual BALSAC ID number more than once per census year is not allowed.

For a given census, an individual can be linked to another census two different ways: a direct link from census to census or a link to BALSAC in the two censuses which allows us to follow the same individual between censuses (see Figure 3). This is the basic material from which it is possible to generate the sequence of individual appearances.

Figure 3. The two types of links which can connect individuals in two censuses



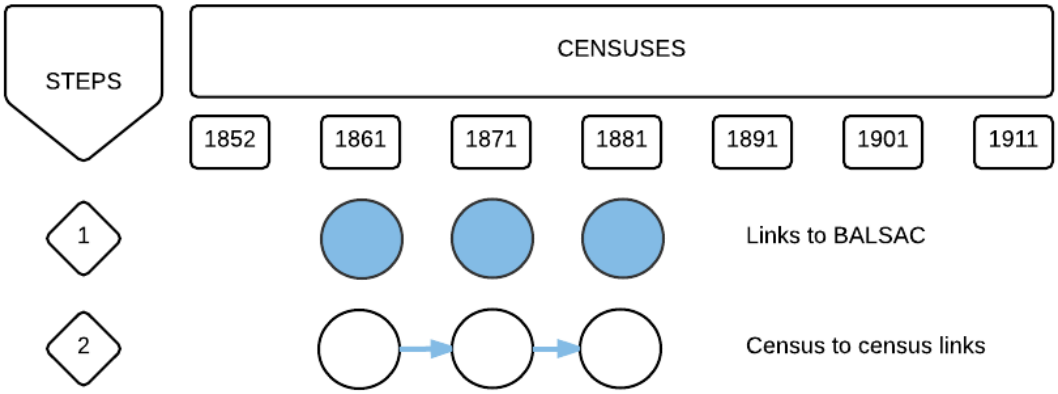
Once the linkage is completed, the next step is to pool the results in order to generate a biographic sequence for each census individual linked to at least one other census. An individual could be linked to BALSAC on all seven censuses and also have been linked to the next census each year which would yield a total of 13 links for a single person. The pros of the method are that the more links are generated, the more we can follow individuals and be confident about their identity. The cons are that more inconsistencies can be created and more validation is needed. But, as we will show, the detection of inconsistencies reinforces the validity of the other links. In order to verify the existence of these inconsistencies and eventually to straighten the results, we must first generate the sequence of census appearances for each individual.

Before describing the process, we will explain why the sequence is generated for individuals only. There are many reasons why we decided not to attempt the same operation on households. First, as mentioned before, once a census individual is linked to BALSAC, we are also able to follow his/her family. A BALSAC family contains all vital events that took place throughout its life cycle and all individuals involved in these events. For that reason, it can sometimes be linked to several households for a given census year. Second, despite being composed most of the time by families like those found in the civil registers, household composition varies over time as people are born and die, move in and move out. For example, say we have a nuclear-type household composed of a couple and their children (household ID number 123). The husband's parents are still alive and living in a different household (household ID number 456). During the period between two censuses, the husband's mother dies. In the next census, we find the father living with his son. In which sequence should we place the household? Should the father's household considered as dissolved? But what if there are still children in the father's household? Can we still consider it as dissolved? One way to address this issue would be to consider the order in which individuals are enumerated in their household, for example by always favoring the married couple over the widowed parent. But that order is not uniform across time, across enumerators or across geographical locations, and any standardization attempt would be time-consuming and costly. Examining various cases, we came to the conclusion that the best way to improve the consistency of the global linkage process is to validate links at the individual level.

Figure 4 shows the steps to create individual sequences. In this example, we follow an individual over three decades as we found him in three consecutive censuses. In the first step we align the census ID numbers linked to the same BALSAC individual. This has the effect of anchoring a unique BALSAC ID number for each linked appearance in censuses. Here, our individual is linked to BALSAC in 1861, 1871 and 1881 censuses. Then, we attach the census to census links made with the second module and as we see a census to census link was made between each pair of censuses. A case like this illustrates a perfect sequence where all links converge to the same unique individual. But individual

sequences are not always as simple: context-related ambiguities are created along the linkage process and distinct identities sometimes overlap from census to census.

Figure 4. Description of the steps to create individual sequences

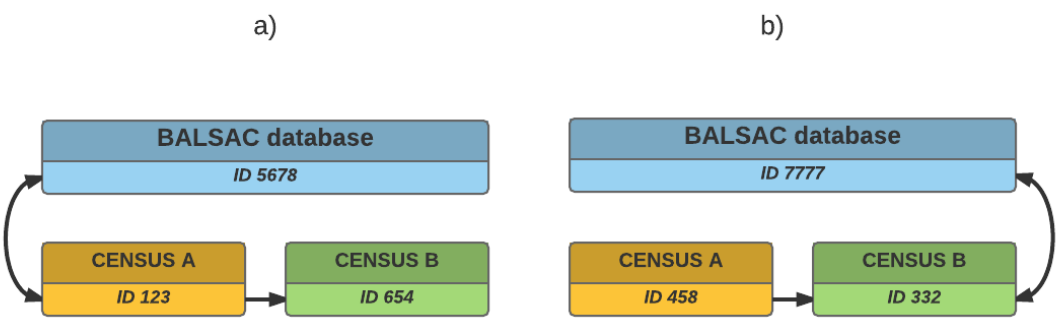


Nevertheless, once generated over the seven censuses, the sequence of links for each individual allows us to perform two closely related operations namely the validation of links and the optimization of linkage results.

Detecting and solving ambiguous links

As mentioned earlier, households from each of the seven censuses are linked to BALSAC families independently using the first module. Next, households from each pair of consecutive censuses are matched independently from the other pairs, with the second module. At the individual level, despite all the household links created, it is possible that some matches will be missed. In each module, the available contextual information is different and thus may be insufficient in one source or another to reach on a decision on the proposed matches. This leads to situations like those presented in Figure 5 where individuals are linked in two consecutive censuses but where a census to BALSAC link was missed in one of the two censuses involved.

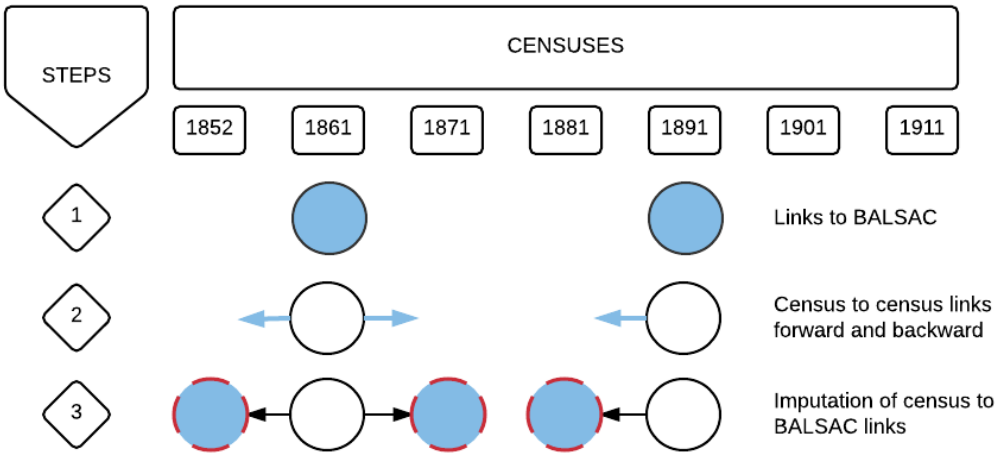
Figure 5. Two examples of ambiguous situations



In terms of linkage results, these are suboptimal situations. They illustrate how the combination of linkage results makes it possible to impute a link which was missed for some reason during linkage operations: from Census B to BALSAC in situation (a) and from Census A to BALSAC in situation (b). In previous work (Bournival et al., 2016), we investigated such specific cases to see if and how it was possible to improve linkage results to BALSAC for the 1891 census using information from the previous and the following census. By triangulation, it was tempting to add the missing link to BALSAC, but we had to make sure that the existing links were accurate in identifying the same individual. The manual work performed to verify these inconsistencies revealed that in only 3 out of 99 ambiguous situations the census to census link was erroneous. There was no instance where the census to BALSAC link was wrong. Based on this result, we made the decision that we could confidently impute the missing census to BALSAC link and consider that if erroneous links were created they would be captured later on in the process.

Figure 6 recalls the steps used to create individual sequences across the seven censuses and shows how missing links can be imputed. We chose an example in which both types of links contribute to the creation of the sequence of an individual found in five consecutive censuses. At each step, the new feature is highlighted in blue and circles represent census to BALSAC links and arrows census to census links. We have to keep in mind the main objective of this operation: we want to maximize links to BALSAC in order to create a fully integrated sequence of appearances for each individual.

Figure 6. Description of the steps to create individual sequences and impute missing links

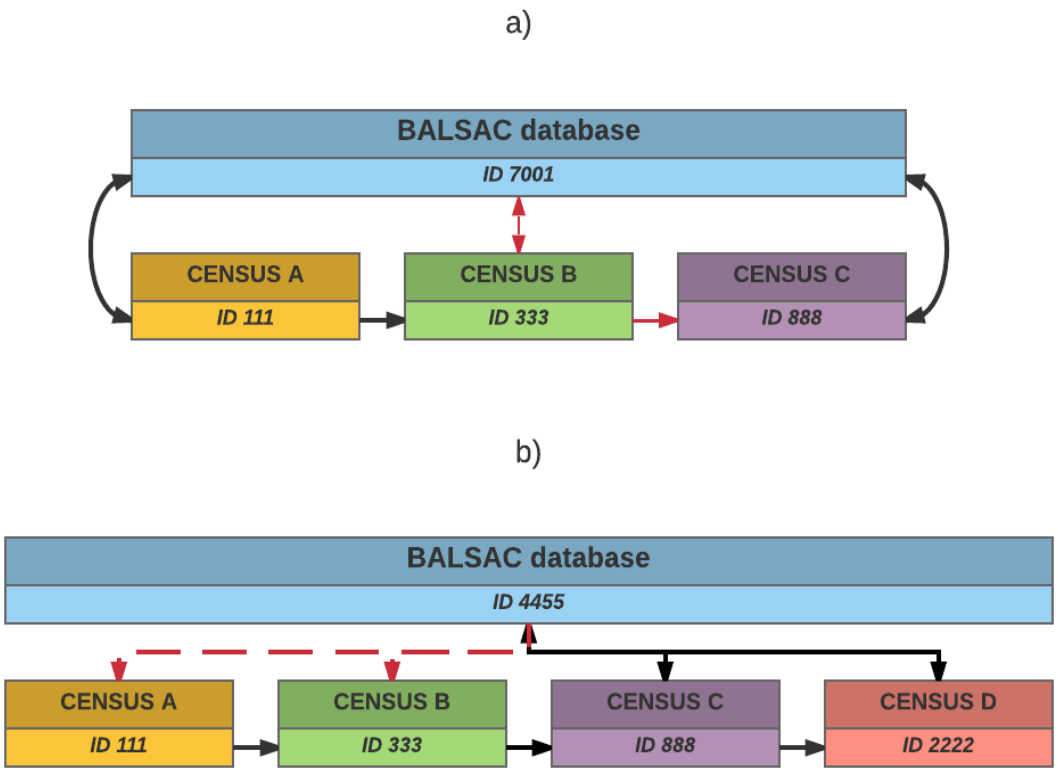


The first two steps are performed exactly as in our previous example (see Figure 4). The main difference here is the presence of gaps in the individual's sequence. Thanks to the pooling of results, it is possible to continue a sequence even when a link was missed with one of the modules. By generating the sequence we see that each census appearance

linked to BALSAC is also linked to one or two other censuses. Following our reasoning, we can retrieve ambiguous situations in 1852-1861, 1861-1871 and 1881-1891 census pairs. The gaps in the sequence can be replaced by implicit census to BALSAC links deducted from the reconstructed sequence.

The logic is the same as above, no matter how many censuses are involved. As an extension of our previous example, Figure 7 illustrates more complex situations involving more censuses in the sequence. The left part of the first diagram is exactly the same as diagram (a) in Figure 5 with a third census added where the individual was found in BALSAC. There is no link between Census B and C but we know that the same individual is identified in BALSAC in A and C and that we identified him in the second year with a census to census link. The red arrows correspond to the links imputed based on these observations. This is maybe more striking in the diagram (b) where an individual is not linked to BALSAC until the third census. Imputation of his or her BALSAC ID number in the two first censuses will restore consistency in what appears to be coherent biographies.

Figure 7. Solving ambiguous situations

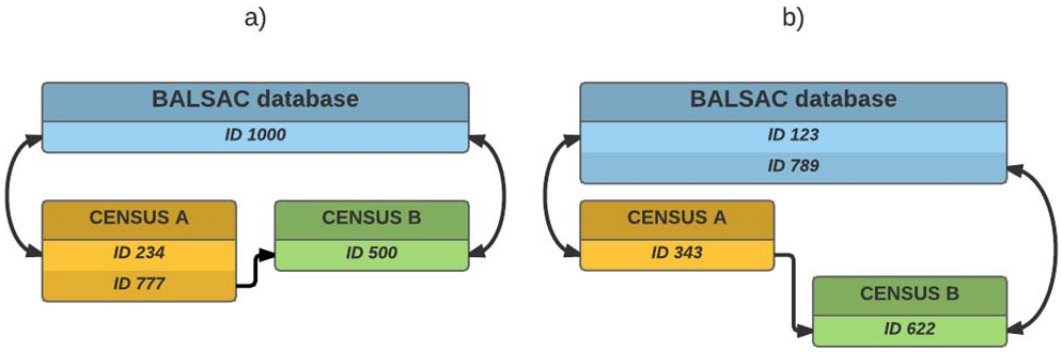


Detecting and correcting erroneous links

Optimizing the links to BALSAC in individual sequences also helps identifying conflictual links across censuses. The use of the sequence as a tool for validation of linkage results becomes more efficient as the number of links involved in the sequence increases. It is important to note that we do not validate what appears to be accurate links that is a coherent sequence of appearances, at least not directly. Instead, we retrieve inconsistencies in census biographies which contain a conflictual link between two distinct individuals, something we call overlapping identities.

First, the sequence detects the links leading to a confusion between two distinct individuals for a given census year. The Figure 8a) represents a situation when an individual was linked to BALSAC in two consecutive censuses, while another individual in the first census was matched to this same individual in the second census via a census to census link. The problem here is that the imputation of the BALSAC ID number to the second individual in census A will lead to a duplication of the BALSAC ID number in the first year.

Figure 8. Two examples of erroneous links



A typical example of a case like the one illustrated in the Figure 8b) would be a daughter (woman A) linked to BALSAC in the first census and matched with her brother's wife (woman B) in the next census as they share similar names and order in both households, and an adequate age gap between censuses. But in the second census the woman B is correctly linked to BALSAC as the wife of woman A's brother. Despite their respective unique BALSAC ID numbers, both women are assimilated to the same individual by the census to census link.

In this particular case, removing the erroneous census to census link will restore consistency in both sequences. In the previous example, we have to validate all the links (three in totals) in order to conclude on the identities of the individuals involved. These are the basic cases of inconsistencies. Other variants exist combining more links, but they all share the same characteristic: they confuse two census individuals.

In theory, correcting the erroneous links for each individual will restore coherence in their respective biographic sequence. But in reality, as biographies are intertwined, it is possible that corrections generate new inconsistencies. What seemed as a coherent census biography might stand out as ambiguous or incorrect once other individual sequences were corrected. Each time a link is modified, suppressed or added, the sequences should be generated again to verify that all the links are still coherent. In the course of this process, even biographies considered as correct are validated.

Results on the Saguenay data

Estimating an error rate and improving linkage rates

Altogether, 48,860 sequences were generated on the Saguenay data and a total 317 erroneous links were detected. Once corrections were done, we performed a second round to verify the presence of new inconsistencies brought about by the corrections made and 23 new errors were found and corrected. After the merging and splitting of sequences due to these corrections, the final number of sequences was 48,734. The ratio between the number of errors and sequences yields an error rate of 0.7%. As our validation approach is not equivalent to a method based on the selection of a random sample of cases to be validated, this rate is difficult to interpret but we believe it globally reflects the quality of data and of the linkage work.

By correcting inconsistencies detected in the pooling of linkage results in individual sequences, we also managed to increase the number of links thus improving not only the quality of the dataset but also the linkage rates both to BALSAC and to the next census. Also, as mentioned before, generating the sequence allows the connection to BALSAC of all census appearances of an individual when there is at least one link to BALSAC in the sequence. In turn, this improves the linkage results between civil registers and censuses. Lastly, correcting the sequence of an individual in a household can lead to linkage of related individuals in the same household if they were not already linked.

Table 2 shows the number of census to BALSAC links added using the individual sequences and the corresponding proportion among individuals who were not linked to BALSAC during linkage operations. As explained above, gains are made through the imputation of links to BALSAC where it was missing in the sequence. We see on Table 2 that these gains are more important between 1871 and 1901 when they represent between 15 and 26% of individuals who could not be linked to BALSAC. A possible explanation could be that individuals benefit from longer sequences (more links to identify and follow them) or, in other words, when the impact of censoring is less important.

Table 2. Links to BALSAC added by generating individual sequences

Census year	Individuals not linked to BALSAC	Links added using the sequence	%
1852	1100	53	4.8
1861	1256	91	7.2
1871	1035	158	15.3
1881	1012	200	19.8
1891	922	241	26.1
1901	1390	212	15.3
1911	2456	147	6.0

Overall, linkage rates to BALSAC for all census years increase by a proportion of 1.2 to 1.6% (data not shown). These proportions appear relatively low, but since the linkage rates for the Saguenay region were already quite high from the start, the possible gains were consequently limited. We believe the increase will be more marked as we work on areas with lower linkage rates.

Investigating the source of errors

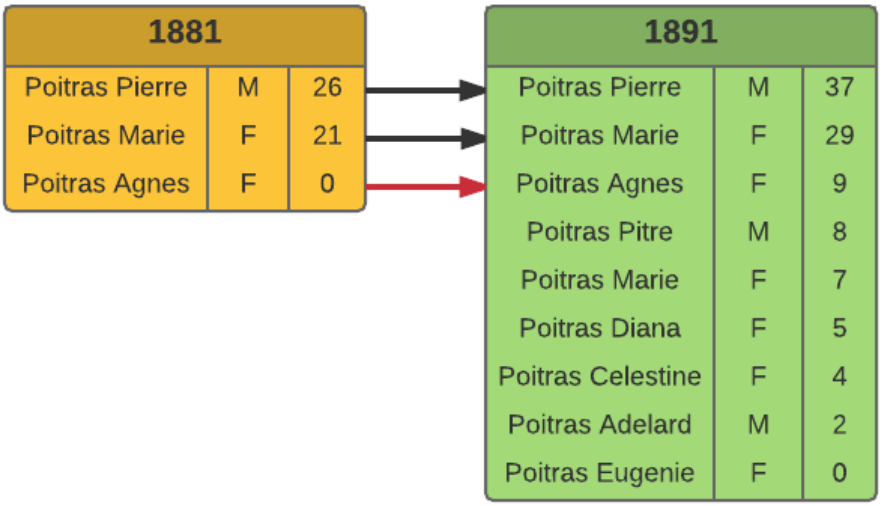
We also attempted to explain the erroneous links by investigating some characteristics of the context in which they occured. Our basic assumption was that most errors would come from the census to census module as mentionned before our trust in census to BALSAC links is stronger based on the available contextual information found in civil registers. It appeared that the difference is not as clear since we found that in 63% of the time inconsistencies came from census to census links meaning that in almost 2 cases out of 5 we had to correct a census to BALSAC link.

For census to BALSAC links, the most common error involves couple homonymy, especially when it happens among childless couples. This indicates that even though it is easier to link to BALSAC because we have at our disposal an already linked family, the homonymy problem remains an important issue. Individuals whose identities were mingled because of homonymy do not appear to share common characteristics as we found errors among couples, children as well as among non-members of the family.

Incorrect census to census links are most of the time due to a lack of context. The most common errors involving a census to census link come from a confusion between sister and sister-in-law or between the two consecutive spouses of a man or between brothers and sisters sharing the same first name. In the census to census module, individuals involved in this type of situation are easily matched if age varies accordingly. For example, we observed several cases where the two consecutive wives of a man shared the same name like “Marie” or even more singular ones like “Adelaide”. Combined with the

surname of the husband, they appeared as the same woman throughout the sequence. Also, it is not so uncommon to link two children sharing the same name when the first one dies and the second one is born during the period between two censuses. Figure 9 displays the real case of Agnes Poitras. In both censuses, Agnes occupies the same rank in the family and her ages are coherent, so there were no reason not to link the two girls. But in the generated sequence, it appeared that both Agnes were in fact distinct individuals. The first Agnes was born in 1880. She was enumerated in 1881 and she died a few months later, in September 1881, while her mother was pregnant. In October of the same year, a second child was born and the parents also gave her the name of Agnes. In the 1891 census, it is the second Agnes who was enumerated. Without the family background provided by the BALSAC database the linkage can only rely on the information found in the next census. That information is not always sufficient and leaves in the shade more complex histories.

Figure 9. A concrete example of overlapping identities



One interesting finding is that despite the quality and the completeness of the civil registers for Saguenay (Bournival et al., 2016; Bourque, Markowski, & Roy, 1984), censuses did help to improve some family files in the BALSAC database. For instance, thanks to the date of birth declared in the 1901 and 1911 censuses, we managed to link birth and marriage records within family files as well as parents and children marriages in about 10 cases where missing information prevented linkage.

Building census-based longitudinal datasets

When generating the sequence, we move from a vision of censuses as a succession of transversal snapshots to a longitudinal dataset. The sequence unfolds a truly longitudinal pattern of linked censuses where we are able to follow a large share of the population over time and space. Each sequence represents a distinct individual that can be followed

as long as he or she appears in the seven censuses[4]. Therefore, these censuses are no longer composed of 94,390 enumerated individuals but of 48,728 distinct individuals as shown in Table 3. Altogether, almost 50% of all individuals appear in two or more censuses. The number of appearances varies according to the period of observation, the timing of individual lives (age, family life cycle) as well as demographic factors like death and migration.

Table 3. Distribution of distinct individuals across the 7 censuses according to their number of appearances

Nb of appearances	Individuals	
	n	%
1	26,212	53.8
2	10,361	21.3
3	5690	11.7
4	3389	6.9
5	1917	3.9
6	910	1.9
7	255	0.5
Total	48,728	100.0

Table 4. Distribution of individuals appearing in only one census

Year	Individuals appearing once		Total census population	% of total census population
	n	%		
1852	1744	6.7	5031	34.7
1861	2273	8.7	8825	25.8
1871	2318	8.8	11,814	19.6
1881	2429	9.3	13,810	17.6
1891	2222	8.5	14,776	15.0
1901	3037	11.6	16,365	18.5
1911	12,189	46.5	23,769	51.3
Total	26,212	100.0	94,390	

We also looked at the census distribution of people who appear only once (see Table 4). This subgroup is composed of unlinked and unprocessed individuals for each census year as well as of mobile individuals who stayed in the region for short periods between two census dates. Moreover, since we have a finite series of censuses, both ends of the spectrum are overrepresented since individuals who died between 1852 and 1861 as well

[4] In the course of this project, two cities and two other regions were also linked to the BALSAC database. Individuals who migrate between these locations in the covered period are of course followed through each of their appearances.

as children born between 1901 and 1911 censuses could not have been enumerated more than once. The 1911 census alone gathers almost half of the individuals of this subgroup.

Concluding remarks

As part of the construction of the IMPQ, we have developed a longitudinal database covering seven censuses and spanning 60 years. This database is linked to families reconstituted using 19th and 20th century Quebec civil records in the BALSAC population database. Linkage was performed at the household and individual levels. It is now completed for the Saguenay region and is ongoing for two cities and two other regions.

In this paper, we have described the final step of our linkage work which involves the pooling of the 13 datasets which contain the linkage results (seven from census to BALSAC linkage operations and 6 six from census to census operations). This pooling is done in order to produce biographic sequences for census individuals. We have shown that generating these sequences, by highlighting inconsistencies, also provides a powerful tool for validating the data, correcting erroneous links and clarifying ambiguous ones.

Overall, our validation process does not prevent the creation of an incorrect link that would fit perfectly in a given census biography as homonymy problems cannot fully be eliminated. But given the high linkage rates, the relatively modest size of the Saguenay population and the numerous consistency checks made on both sources, we can conclude that major errors are very unlikely and that the vast majority of the created links are accurate.

Following linkage operations, we have found that in 75% of cases, links are created with the candidate considered the best one by the program (candidate with the highest score). Based on this observation and knowing that we could rely on individual sequences to detect inconsistencies, we are currently considering the possibility of introducing an automatic linkage option in our program. Since we intend to pursue the development of the IMPQ in the coming years and we aim at expanding the type of data to be linked, this would represent a valuable improvement.

To conclude, researchers involved in the construction and linkage of population databases are concerned with issues of performance of their approach in terms of quantity (linkage rates, completeness) and quality (accurateness and representativeness of the links). In both instances, validation is an important but challenging component of the linkage process as the possibilities are limited and highly dependent on the method used. The work presented in this paper suggests that when more than one data source is used the combination and comparison of linkage results obtained from each of these sources can provide a helpful contribution to the validation strategy. More widely, although data and methods are highly variable, it could be possible to work collectively and propose some basic guidelines aimed at the development of validation tools in population data linkage.

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 Tremblay (Université du Québec à Chicoutimi). We also acknowledge the support of research professionals, clerks and technical staff involved in this project.

References

Antonie, L., Inwood, K., Lizotte, D. J., & Andrew Ross, J. (2013). Tracking people over time in 19th century Canada for longitudinal analysis. *Machine Learning*, *95*(1), 129–146. http://doi.org/10.1007/s10994-013-5421-0

Bouchard, G., Brard, P., & Lavoie, Y. (1981). FONEM: Un code de transcription phonétique pour la reconstitution automatique des familles saguenayennes. *Population (French Edition)*, *36*(6), 1085–1103. http://doi.org/10.2307/1532326

Bouchard, G., & Larouche, J. (1988). Dynamique des populations locales: la formation des paroisses rurales au Saguenay (1840-1911). *Revue D'histoire de l'Amérique Française*, *41*(3), 363–388. http://doi.org/10.7202/304583ar

Bouchard, G., & Pouyez, C. (1980). Name Variations And Computerized Record Linkage. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *13*(2), 119–125. http://doi.org/10.1080/01615440.1980.10594037

Bouchard, G., Roy, R., & Casgrain, B. (1985). *Reconstitution automatique des familles: le système SOREP*. Chicoutimi: SOREP.

Bournival, J.-S., St-Hilaire, M., & Vézina, H. (2016). Comparing information from vital events to the 1891 census data in the Saguenay region of Quebec: a critical appraisal of the two sources.

Bourque, M., Markowski, F., & Roy, R. (1984). Évaluation du contenu des registres de l'état civil saguenayen, 1842-1951. *Archives*, *16*(3), 16–39.

Fleury, M., & Henry, L. (1956). *Des registres paroissiaux à l'histoire de la population: manuel de dépouillement et d'exploitation de l'état civil ancien*. book, Paris: I.N.E.D. Retrieved from http://ariane.ulaval.ca/cgi-bin/recherche.cgi?qu=a2188992

Goeken, R., Huynh, L., Lynch, T. A., & Vick, R. (2011). New Methods of Census Record Linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *44*(1), 7–14. article.

http://doi.org/10.1080/01615440.2010.517152

Ruggles, S. (2002). Linking Historical Censuses: a New Approach. *History and Computing*, *14*(1–2), 213–224. http://doi.org/10.3366/hac.2002.14.1-2.213

Vézina, H., St-Hilaire, M., & Bellavance, C. (2014). The linkage of micro census data to vital records: New perspectives for longitudinal studies. In *Using micro census data: recent developments and new perspectives*. First Conference of the European Society of Historical Demography (ESHD). Alghero, Sardinia, Italy. September 25th-28th.

Vézina, H., St-Hilaire, M., & Bournival, J.-S. (2015). The linkage of micro census data and vital records: an assessment of results on Quebec historical censuses (1852-1911). Paper presented at the 40th Annual meeting of the Social Science History Association, Baltimore, MD, November 12-15.

Vick, R., & Huynh, L. (2011). The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *44*(1), 15–24. article. http://doi.org/10.1080/01615440.2010.514849

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing Methods of Record Linkage on Swedish Censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *47*(3), 138–151. article. http://doi.org/10.1080/01615440.2014.913967