# Using Collaborative Genealogy Data to Study Migration: a Research Note

Arthur Charpentier, Ewen Gallic

# Using Collaborative Genealogy Data to Study Migration: a Research Note*

Arthur Charpentier[a] and Ewen Gallic[†b]

[a]Université du Québec à Montréal (UQAM), Quantact & CREM UMR CNRS 6211.
[b]Aix-Marseille Univ., CNRS, EHESS, Centrale Marseille, AMSE.

January 15, 2019

**Abstract**

The digital age allows data collection to be done on a large scale and at low cost. This is the case of genealogy trees, which flourish on numerous digital platforms thanks to the collaboration of a mass of individuals wishing to trace their origins and share them with other users. The family trees constituted in this way contain information on the links between individuals and their ancestors, which can be used in historical demography, and more particularly to study migration phenomena. The case of 19th century France is taken as an example, using data from the family trees of 238,009 users of the Geneanet website, or 2.5 million (unique) individuals. Using the geographical coordinates of the birthplaces of 25,485 ancestors born in France between 1800 and 1804 and those of their descendants (24,516 children, 29,715 grandchildren and 62,165 great-grandchildren), we study migration between generations at several geographical scales. We start with a broad scale that of the departments, to reach a much finer one, that of the cities. Our results are consistent with those of the literature traditionally based on the parish or civil status registers. The results show that the use of collaborative genealogy data not only makes it possible to support previous findings of the literature, but also to enrich them.

**Keywords** : Genealogy; Collaborative data; Migration; 19th Century

*This research was conducted within the 'ACTINFO' project under the aegis of the Risk Foundation, a joint initiative by the GENES, the University of Rennes 1, the University of Paris-Est La Vallée and COVEA. A preliminary work was presented during the 'Science XXL' days organized in March 2017 at the French Institute for Demographic Studies (INED). We thank Olivier Cabrignac and Jérôme Galichon for their help on data exploration, as well as the participants for the discussions we had then, which motivated some of the elements presented in this study. We also thank the members of INED's History and Populations Unit for their comments. We benefited from fruitful discussions with the participants of the 'UseR' conferences (Budapest, May 2018), the 'R meeting' (Rennes, July 2018), the 'XXIX International Biometric Conference' (Barcelona, July 2018), and the 'Eco-lunch' seminar (Marseille, September 2018)

[†]Corresponding author: ewen.gallic@gmail.com. Aix-Marseille School of Economics, Aix-Marseille Université, 5-9 Boulevard Bourdet, CS 50498, 13205 Marseille Cedex 1, France

# 1 Introduction

Historical demography is a discipline founded in the second half of the 20th century aimed at studying the population in the past. The pioneering work of Henry (1956) has given rise to a multitude of studies on topics such as mortality (Blayo and Henry, 1967; Blayo, 1975a; Henry and Blayo, 1975) and family structure (Matthijs and Moreels, 2010). Migration has also been studied, although neglected in population history and in family history, as Oris (2003) rightly points out. Although the population of pre-modern societies has not been static, migration increased after the middle of the 19th century, mainly due to improved transport, according to Lucassen and Lucassen (2009). Other factors have been put forward to explain migrations. They can be divided into two categories: the first involves the role of the family, and the second explains migration through social background and education.

Regarding the role of the family in migration decisions, Kesztenbaum (2008) showed the positive influence of siblings: having a brother who had already migrated has a positive impact on the migration of other brothers. However, the author notes that network effects play only a weak role in destination choices. This poor network effect is also reported by Bonneuil et al. (2008). Family composition also plays an important role. Dribe (2003) showed that the number of elderly individuals has a negative influence on the propensity to migrate, while conversely, the number of young individuals plays a positive role in the propensity to migrate.

Social environment and education are two other explanatory factors of migration highlighted in the literature. The research of Heffernan (1989) has exposed the existence of a positive correlation between literacy and migration. This link was examined by Bourdieu et al. (2000). These authors have shown that the higher the level of education, the more likely individuals are to migrate. Rosental (2004) and Bonneuil et al. (2008) obtained similar results. Rosental (2004) also pointed to a distinction between urban and rural areas in the 19th century, with urban populations tending to be more attracted to cities and more willing to travel long distances.

Many empirical studies in historical demography, whether or not specifically concerned with migration phenomena, rely on register data. The information provided by the registers relates to the dates of births, marriages and deaths of individuals. Louis Henry's interest in such data as early as the mid-1950s is at the root of historical demography. His conceptual framework coupled with the use of parish registers and the construction of a family form have attracted the attention of his peers and have contributed to the success and enthusiasm for the discipline (see Rosental and Mandelbaum, 2003 for more details). However, this new approach towards demography had some limitations. As noted by Chamoux (1972), reconstituting families using register data is a tedious task, and generalizing results obtained from small samples can be a challenging one. However, it should be noted that the TRA project, initiated by Jacques Dupâquier, which relies on the reconstruction of families based on register data, provides a representative sample of the French population (see Bourdieu et al., 2014).

In France, although parish and then civil registers since the French Revolution have been widely used, they are not the only source of information in the literature. For example, military registration numbers established by the military administration constitute a different source, used in particular to study migration during life (Ho., 1971; Kesztenbaum, 2008, 2014). A source closely linked to the existence of the various reg-

isters, since it relies on them, has contributed to the improvement of knowledge of the history of our ancestors: genealogical data. It is noteworthy that Mormons have built up extensive genealogical databases at the University of Utah that have contributed to population studies in the past (Bean et al., 1978; Lindahl-Jacobsen et al., 2013). Recently, with the era of the development of computer techniques, this type of data seems to benefit from an increasing interest. The digital revolution makes it possible to access genealogical information in a relatively short period of time, at a lower cost and with less effort. Many websites offer their users to reconstitute their family trees. This is the case of wikitree.com, familysearch.org or geni.com. So far, the information recorded by the users has mainly been used to study the longevity of individuals (Gavrilova and Gavrilov, 2007; Gavrilov et al., 2002; Cummins, 2017; Fire and Elovici, 2015).

Recently, Störmer et al. (2017) used crowd-sourcing genealogical data to study marriage timing and fertility levels in the Netherlands. Their results suggest that migration may have played a key role in explaining an increase in the age at marriage as well as a lower fertility in Europe between the 17th and 19th centuries. Kaplanis et al. (2018) also used collaborative genalogy data, to explore the family trees of several million individuals. Their study shows that genealogical data obtained through the collaboration of amateurs can produce high quality genealogical trees. We use the same type of data to study migrations, with an application to the French case in the 19th century. More specifically, we exploit information provided by amateur and professional genealogists who have built their family tree on a website called Geneanet. Users of the site have given information about the places and dates of birth, marriage and death of their ancestors. Based on this information, we study the migration of descendants of individuals born at the beginning of the 19th century in France. The concept of migration can be understood in different ways, as recalled by Greenwood (1997). One of them, proposed by the United Nations, states: '*A migration is defined as a move from one migration defining area to another (or a move of some specified minimum distance) that was made during a given migration interval and that involved a change of residence*' (United Nations, 1970). This definition involves two notions: a temporal one, and a spatial one. This distinction can be found in the classification into four types proposed by Fine (1991, pp. 88–89). In his opinion, the four types of migration can be classified according to their relationship to time, ranging from short-term to long-term. The first type of migration concerns commuting between home and work, which is carried out periodically. The second type of migration characterizes seasonal movements related to economic activity, such as agricultural harvesting. In this case, individuals migrate temporarily so as to stay close to places where economic activity has temporarily increased during the season. When the latter ends, labor supply decreases and seasonal workers return home. The third category of migration is also temporary, but lasts longer. It describes temporary movements during a lifetime in connection with a change in activity. These first three types of migration are closely linked to the labor market.[1] The fourth type of migration is broader. It describes the permanent movements of individuals from the mountains to the countryside, and from the countryside to the cities. This study focuses on this fourth type of migration due to the data it uses.

This article contributes to the literature in historical demography through a study of

[1]The labor market is used to define a migrant. Shryock and Siegel (1976, p. 374) refer to the existence of increasing opportunity costs of distance from home to work, which can lead to a change of residence when it becomes too high.

the French case in the 19th century. It presents observations on migration from generation to generation, drawing on rich individual data from the collaboration of hundreds of thousands of Internet users. The internal migration of the French is described through the prism of several spatial scales, ranging from the global level to a much finer level. The results show that the use of collaborative genealogy data not only makes it possible to recover known facts in the literature, but also to enrich them.

The rest of this article is organized as follows. Section 2 presents the data. Section 3 examines migration at the national level. Section 4 takes a more detailed look at the types of migration by distance travelled. Section 5 proposes to examine movements between cities according to city size. Finally, Section 6 addresses the special case of migration to Paris, the capital of France.

## 2 Data

The analysis relies on collaborative data from a genealogy website, Geneanet.[2] On that website, users looking for their ancestors collect information themselves, and fulfill their family tree.[3] The information they provide about their ancestors corresponds to three types of events that can be found on civil and religious registers: birth, marriage, if any, and death. For each event, a date and a place can be fulfilled. Individuals in a tree are linked through their parents and spouses.

In this article, we focus on French migration in the nineteenth century. We follow the movements of people who were born in France[4] between 1800 and 1804,[5] and their offspring over three generations. Raw data includes $701,466,921$ observations from the trees of $238,009$ distinct users. A raw observation contains information about one or more events in the life of an ancestor, within a user's family tree. These events are, as previously mentioned birth, possible marriage(s), and death. When the events of the same person in a user's tree take place in the same geographical location, they are grouped into a single observation. A substantial task of matching and data cleaning is, however, needed.[6] First, as each amateur genealogist constructs their family tree on the website, there are a lot of duplicated individuals in the raw data. We need to match individuals referring to a single person. Second, we need to clean the data to: $(i)$ complete missing observations during the merger of trees, and $(ii)$ correct some obvious mistakes. Completion of information can be done thanks to the large amount of data. Let us consider an example in which some information is missing, say the place of birth of an individual in a user's tree, but other information is present, *e.g.*, the date of death. In the family tree of another user, the record that refers to the same individual provides details on the place of birth but lacks information regarding the date of death. Then, during the merger, we are able to complete the life events of the individual. The correction of mistakes also happens during the merger. If a user misspelled the first name of an individual, it is possible to correct it, provided that a majority of other users agree on another spelling. The

[2] https://www.geneanet.org/

[3] The website's users can choose between publicly sharing their family tree or keep it private. We do not have access to the latter, so that our analysis only uses publicly shared family trees.

[4] According to Geneanet, 40% of their records relate to French data.

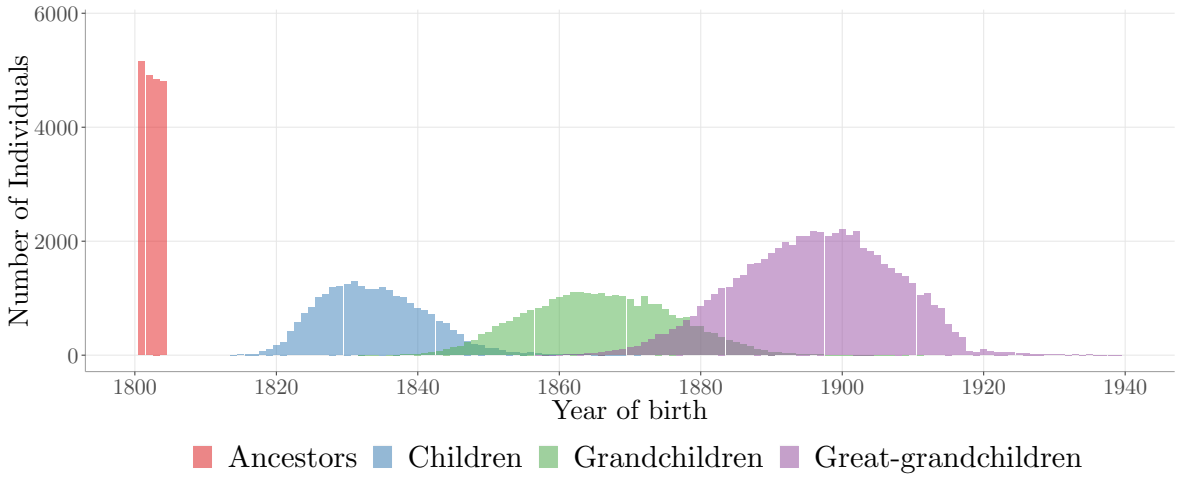
[5] As reminded by Fleury and Henry (1958), this period corresponds to the uninterrupted recovery of death, marriage and death records in the whole of France.

[6] More details can be found in Appendix A.

complete methodology as well as the computer codes (written in R) are provided in an online annex available at the following address: http://egallic.fr/Recherche/Genealogy/.

Once the family trees are matched and the data is cleaned, we proceed to a restriction on age. As we are interested in migration, we only consider people who are in a position to move independently and according to their own rationale that is not the one of their parents. We thus arbitrarily keep people who survived at least until they are 16 years old and discard the rest of observations.[7] We end up with 25,485 ancestors born between 1800 and 1804, 24,516 children, 29,715 grandchildren and 62,165 great-grandchildren[8]. The distribution of the year of birth, by generation, is shown in Figure 1.

Figure 1: Distribution of Year of Birth in the Sample, by Generation.




Note: This graph shows the distribution of birth years for individuals born in metropolitan France between 1800 and 1804, and for their descendants over three generations (blue for children, turquoise for grandchildren, and mauve for great-grandchildren). The number of births per year is indicated on the y-axis, using a logarithmic scale.

# 3 Migration at the National Level

To get a global idea of regional heterogeneity of internal migrations, we rely on the information given regarding the place of birth of the individuals who were born in France between 1800 and 1804 and that of their descendants. More specifically, we extract the administrative area in which these individuals were born. These areas, which are called department, were established in 1790. They divide the European territory of France in 96 pieces.[9] Once we have obtained the birth department for both ancestors and their offspring, we compute the percentage of descendants for each ancestor, in each

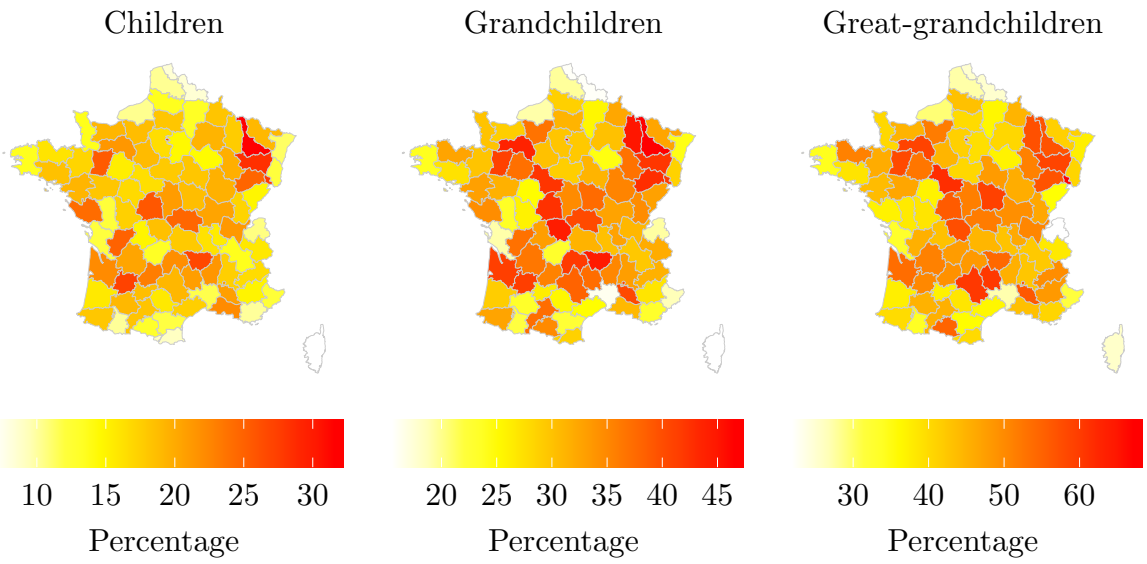
[7] There is evidence of child labour at younger ages (Fauve-Chamoux, 2004), which may suggest that a lower threshold could be established. Lowering the threshold to 14 years of age only changes the results at the margin and leads to the deletion of 451 observations..

[8] As the objective is to observe internal migration from one generation to the next, we consider each ancestor born between 1800 and 1804 in metropolitan France and follow all his or her descendants present in the data. Thus, the child of a couple will be present twice in the data if information on their parents is available. It should be noted, however, that the numbers of 24,516 children, 29,715 grandchildren and 62,165 great-grandchildren refer to unique individuals in the database.

[9] Currently, there are 101 department in France, 96 of which are located in the European territory, so-called 'metropolitan France'.

department, who were born in another administrative area. The resulting percentages are shown in Figure 2, for each generation. The majority of the children of the individuals that make up the sample of ancestors were born in the same department as their ancestors. A smaller proportion is observed for the grandchildren and an even smaller one for the great-grandchildren. Significant regional differences can be seen. There are indeed proportions of descendants to be born in a department different from that of their ancestor relatively higher in the center of the country compared to the rest of France.

Figure 2: Percentage of Descendants Born in a Department Different from that of their Ancestor, by Department.




Note: These maps show the percentage of children (left), grandchildren (middle) and great-grandchildren (right) that were born in a department different from that of their ancestor who was born in France between 1800 and 1804. The percentages are represented on a spectrum that ranges from yellow (low values) to red (high values), specific for each map.

It is possible to focus only on the descendants who were born in a different department from that of their ancestor. For these movers, we extract the coordinate pairs (longitude, latitude) and estimate the density of births over the French territory, for each generation and each department. We use a 'modified' Gaussian Kernel to estimate these densities.[10] We represent them using heat maps. An example is provided in Figure 3, for two French administrative divisions 'Indre-et-Loire' (Figure 3a) and 'Lot' (Figure 3b) which are located in the center-west and southwestern France, respectively. For each map, the corresponding birth department of the ancestors are filled with gray. As can be seen in both examples, most of the descendants were born not far from the department in which their ancestor was born. The Paris region appears to be a common birthplace for individuals born outside the department of origin of their ancestors, except for descendants of the first generation.

Apart from these common points, the differences observed on the maps of these two

[10] As we are interested in internal migrations in France, we only focus on French territory. In addition, we are looking at the places of birth of the descendants who were born outside the department of their ancestor. So, when computing the density of births for individuals born outside a given department, we know *a priori* that the birthplace of these individuals cannot be found inside that department. This information should be accounted for when estimating the density of births, to avoid a border bias. The procedure is explained in Charpentier and Gallic (2016).

departments suggest different migration strategies. Indre-et-Loire is marked by a low number of births outside the department. The Lot region highlights a relatively more accentuated migration.

The department of Indre-et-Loire corresponds to the former province of Touraine, including the eastern part of the former province of Anjou. Its main city is Tours, populated by 20,240 inhabitants in 1800 and 63,267 in 1900. At the beginning of the 19th century, Tours was still a city enclosed by its 17th-century walls. The city of Tours is not yet overflowing, as the entire south-eastern part is not urbanized. Outside the ramparts, the neighbouring communes also remain essentially rural, with the Varennes still dominating the landscape.

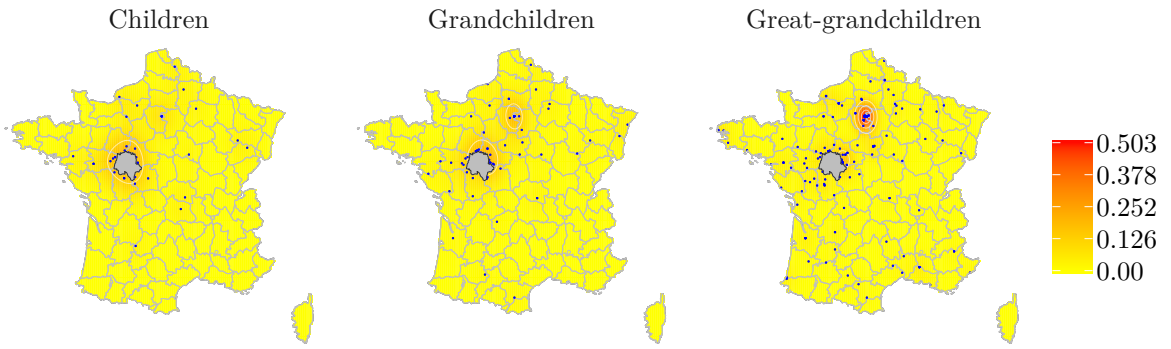
The department of the second example, the Lot, has a different profile. Its main city, Cahors was populated with 11,728 in 1800 and 14,502 in 1900. The Lot department was formed from the eastern part of the province that was named Quercy before 1789. The traditional Pyrenean family system, as a principle of social organization, is one of the most historically characteristic examples in the world of house-to-house system, a concept of social sciences now assimilated to that of the ancestral family described by Frédéric Le Play on the basis of the Pyrenean system. This system of customary law, which has also been described more widely (in more or less attenuated forms) in south-western France and northern Spain, was distinguished in particular by a single succession, on the contrary, for example, of related systems (division of assets among all heirs). This family system of heritage conservation, in which the main aim is to keep possession of agricultural land and livestock, or even to make them grow, therefore favours a low birth rate (too many cadets are seen as a social burden), marriages arranged according to economic imperatives (arrangement between clans or marriages within the clan), and the emigration of cadets in order to avoid additional social costs. As cadets in general did not inherit, it was quite common to emigrate, as discussed in Pinède (1954): there was a strong immigration to South America (that we do not observe in our data). And as mentioned in Fourastié (1986), some villages were among the most closed isolates known in France, with a very high rate of endogamous marriages. What we see on the maps echoes these observations. Many hotspots can be observed, such as the one in northern France. The mining operations of the region and their labour requirements can be highlighted to explain the choice of this emigration spot. Another economic pole stands out: the region of Marseille, for the second and third generation (grandchildren and great-grandchildren, respectively).

# 4 Short and Long Distance Migrations

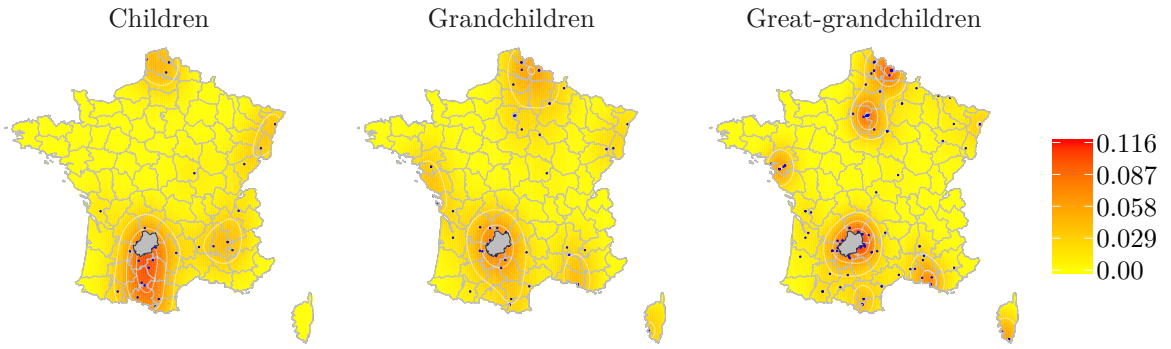
Rather than considering migration by the prism of regional changes, it is possible to leave the regional borders aside and to consider only the distances travelled from one generation to next. By using the geographic coordinates associated with the places of birth of the individuals in our database, it is possible to calculate the distances separating the places of birth of an individual from those of his or her descendants. We do it for the ancestors who were born in France between 1800 and their descendants.[11] Contrary to what we did for the analysis of migration between departments, we consider births outside the

[11] See more details in Appendix B regarding the methodology used to compute distances.

Figure 3: Birthplace of Descendants Born Outside the Birth Department of their Ancestor.




(a) Indre-et-Loire




(b) Lot

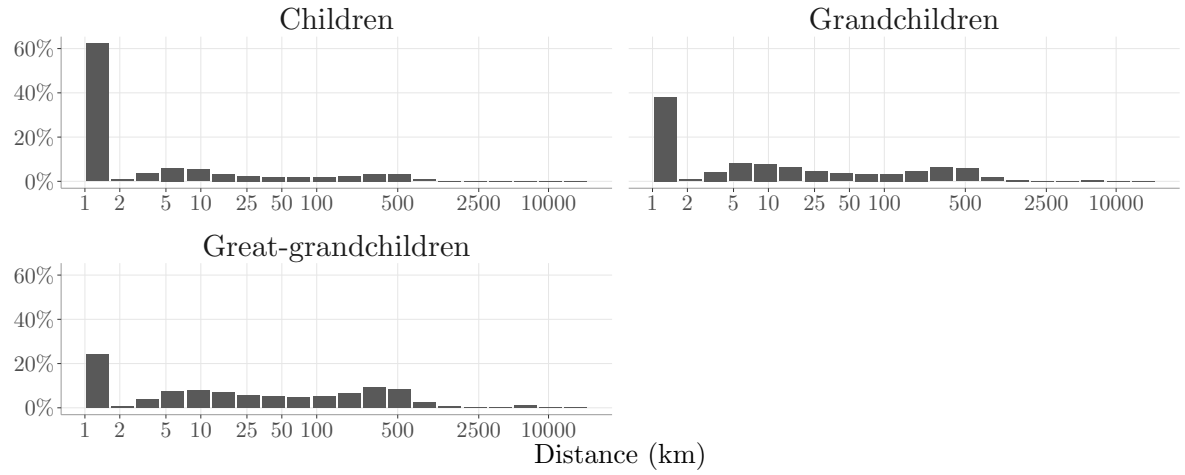
Note: These maps show the estimation of the density of the birthplace of the descendants of individuals born between 1800 and 1805 in two different French departments: Indre-et-Loire (top) and Lot (bottom). We excluded the descendants who were born in the same department as their ancestor. The density estimations are represented on a spectrum that ranges from yellow (low density) to red (higher density). Each place of birth is represented by a blue dot.

territory of Metropolitan France.[12] The distributions of distances between generations are provided in Figure 4. The graph shows that a large part of the descendants, regardless of their generation, were born at exactly the same place as their ancestors. Their share, however, decreases gradually over the generations, from 60% for children to just over 20% for great-grandchildren.

Like the results put forward by Bourdieu et al. (2000), we observe bimodality in the distribution of distances among those who were not born in the same place as their ancestor. We can use this bimodality to disentangle between the short distance and the long-distance migrants. It is common in the literature to make such a distinction, although there is no consensus regarding the value that allows one to differentiate short-

[12]However, the number of descendants born outside France in the data is low, with only 261 observations.

Figure 4: Migration Between Generations.




Note: The graphs show the distribution of birth distances between the birthplace of ancestors born in metropolitan France between 1800 and 1804 and that of their descendants (children, grandchildren, great-grandchildren). The axis of the y-axis indicates the percentage of descendants whose place of birth is located at a specific distance (indicated on the y-axis) from the place of birth of their ancestor. Please note that the distance is plotted on a logarithmic scale.

distance migrants from long-distance ones. While Rosental (2004) uses a value of 25 km, Bourdieu et al. (2000) use 20 km and Kesztenbaum (2008) places the cursor at 17 km. All of them choose their value according to their data, so that it reflects the median distance travelled by migrants. In our sample, the median value for the individuals from the first generation, *i.e.*, the children, is 20 km. We therefore use 20 km as the threshold value for separating short-distance and long-distance migrants. The idea behind separating migrants according to the distance they travel is based on economic and sociological arguments. As explained by Kesztenbaum (2008), the higher the distance, the higher the costs, whether economic or not. Rosental (2006) argues that the social status of individuals is one of the factors that explain why people decide to move to another place. According to him, the most privileged, who were also those who received a better education, tend to travel long distances while people from modest background tend to travel short distance, and sedentariness represents an average.

In our sample, there is a strong sedentariness for our ancestors, which amounts to 62.17% (Table 1). We note that fewer of their grandchildren are born in the same place (38.06%), and that for great-grandchildren, only 24.17% of them are born in the same village as their grandparents. At the same time, it should be noted that the proportion of descendants born at a long distance from the ancestors' birthplace has increased relatively more than that for short distances.

A gender perspective on migration can also be provided using collaborative genealogy data, although as Rosental (2004) points out, gender is the second factor to look at in order to differentiate migration, the first being more of an educational issue (unfortunately not available with current data). To this end, we look at the different migration patterns of ancestors (*i.e.*, individuals born in metropolitan France between 1800 and 1804) according to their gender. We keep the same distinction as before, by defining three categories to describe migration: born in the same place, at a short distance, and a long distance from the birthplace of the ancestor.[13] The results are also shown in Table 1.

[13]It should be noted that people who had no children are excluded from this analysis.

Table 1: Proportions of Descendants Born in the Same Place, at Short or Long Distance from the Birthplace of their Ancestors.

| | Children | | | Grandchildren | | | Great-grandchildren | | |
|---|---|---|---|---|---|---|---|---|---|
| | Female | Male | All | Female | Male | All | Female | Male | All |
| Same place | 61.88 | **66.07** | 62.17 | 38.05 | **41.78** | 38.06 | 24.32 | **27.27** | 24.17 |
| Short distance | **22.14** | 14.6 | 19.43 | **29.48** | 24.15 | 27.70 | **28.22** | 24.8 | 27.06 |
| Long-distance | 15.97 | **19.33** | 18.40 | 32.47 | **34.07** | 34.24 | 47.46 | **47.93** | 48.77 |
| Total | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

Note: This table shows the percentage of descendants born in the same place, at short distance ($\leq 20km$) and long-distance ($> 20km$) from their ancestor's birthplace, for each generation, according to the gender of the ancestor (in columns). The migration distance used corresponds to the number of kilometers separating an individual's birthplace from that of his or her ancestor. The values in bold indicate which proportion of descendants is higher between those of women and men.

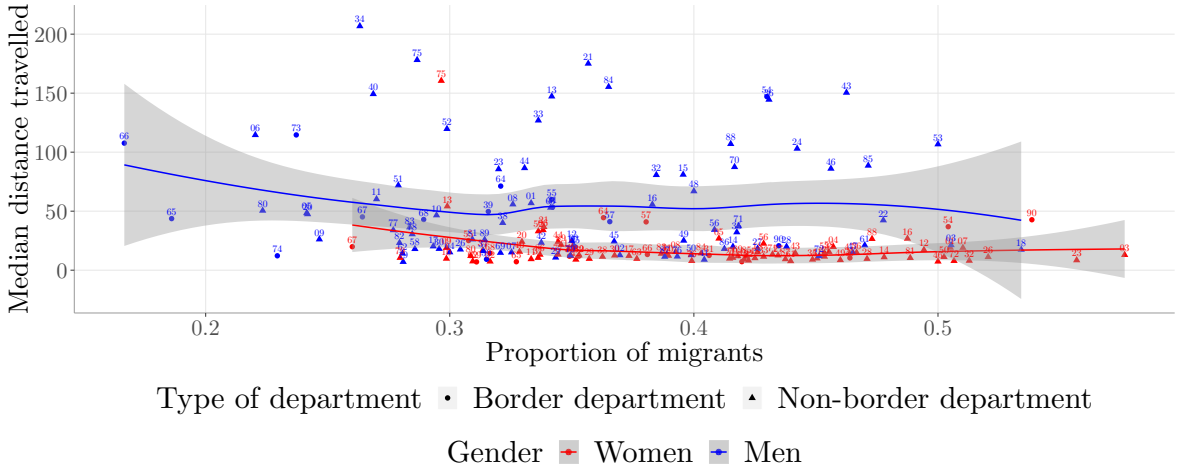
As can be seen, women tend to migrate more but not as far as men, which is consistent with the results provided by Ravenstein (1885) and recalled by Rosental (2006). It can be noted that about 62% of women born between 1800 and 1804 had children in the same place as their own birthplace, which is (significantly) lower than the proportion for men. The bold values in the table indicate which proportions are larger between those of women and men. It can be noted that the pattern showing that tend to migrate more but less far away is repeated over generations. However, while there is a significant immediate effect of gender, it tends to decline over generations, as the differences in proportions tend to narrow between women and men.

The same type of results can be observed at the departmental level, as shown in Figure 5. The latter represents the median migration distance per department (y-axis) as a function of the proportion of migrants for that department.The migration distance is measured this time as the difference, for each individual, between his or her place of birth and that of his or her first child. Each point represents this pair of values for a department, whose official numbered code is indicated above. We distinguished the values for women (in red) from those for men (in blue). A loess smoothing allows us to observe several things. First, regardless of gender, there is a slightly decreasing relationship between the distance of migration and the proportion of migrants. Thus, the higher the proportion of migrants in a department, the shorter the distances travelled. The gender distinction also shows that, as observed at the national level, women tend to migrate more than men, but not as far as men. However, thanks to the different shape of the points, we note that the proportion of migrants for border departments (represented by triangles) is lower than that of non-border departments (represented by circles). This difference may be due to the fact that the trace of individuals emigrating to a different country is more difficult for the genealogist to find, and therefore does not appear in the data.

# 5 Migration Between Municipalities

A look at intergenerational migration can also be done at the municipality level. We consider that the place of birth of a newborn corresponds to the place where his or her parents are living. With recent data, this hypothesis would lead to important biases, as the municipalities in which the newborns are registered correspond to those of the maternities in which the newborns are born. However, for the period considered, *i.e.*, the nineteenth century, a vast majority of birth deliveries take place at home. As reminded by (Fine, 1991, p. 34), 98% of deliveries still take place at home by the eve of the Second

Figure 5: Median Distance Travelled as a Function of the Proportion of Migrants in Each Departement..




Note: Each point represents the median migration distance of individuals in a department (y-axis) as a function of the proportion of migrants for that department (x-axis). The numbers above the points indicate the code of the corresponding departments (as defined by the official geographical code). The red dots indicate the values for women, the blue dots indicate the values for men. The shape of the points distinguishes between border departments (triangles) and non-border departments (circles). The lines accompanied by a confidence interval are obtained by means of a loess regression. Migration distances are obtained by comparing an individual's place of birth with that of his or her first child.

World War.

Information on population size of municipalities in terms of inhabitants can be added to classify the municipalities in different groups. Values for the year 1838 are provided at the scale of chief town and cities (Statistique Générale de la France, 2010). We distinguish between four categories, as in Fleury and Henry (1958) and Blayo (1975b):

- large cities: more than 50,000 inhabitants ;
- medium cities: between 10,000 and 50,000 inhabitants ;
- small cities: less than 10,000 inhabitants ;
- villages: municipalities that are not included in the *Statistique Générale de la France* data.

The number of large cities is 9: Paris, Lyon, Marseille, Bordeaux, Rouen, Toulouse, Nantes, Lille, and Strasbourg. The number of medium and small cities is 98 and 254, respectively. The remaining municipalities in our sample is classified as villages. A vast majority of births occur in the latter, as shown in Table 2. For the ancestors, *i.e.*, those who were born in France between 1800 and 1804, 88% of them were born in a village. Their descendants were also mainly born in villages. It can, however, be noted that the share slowly declines until 80% by the great-grandchildren. In the meantime, the share of individuals that were born in a large or a medium city grows from 3.40% for the ancestors to 7.63% for the great-grandchildren. A similar growth is observed for medium cities. On the contrary, the share of individuals in each generation that were born in a small city remains almost unchanged.

The decline observed in the proportion of grandchildren that were born in a small city can be linked to the phenomenon of rural exodus, which begins in the second half of the nineteenth century, according to Lemercier and Rosental (2000).

Table 2: Distribution of the Places of Birth of the Individuals in the Sample According to the Size of the City.

| Type of Municipality | Ancestors | Children | Grandchildren | Great-grandchildren |
|---|---|---|---|---|
| Large | 3.40 | 3.92 | 6.31 | 7.63 |
| Medium | 5.18 | 5.25 | 6.55 | 7.86 |
| Small | 3.43 | 3.56 | 3.73 | 3.81 |
| Village | 87.99 | 87.27 | 83.41 | 80.70 |
| Total | 100 | 100 | 100 | 100 |

Note: This table shows the proportion of individuals in the sample for each generation (ancestors, children, grandchildren and great-grandchildren, in column), that were born in one of the four categories of municipalities: large city (Large), medium-sized city (Medium), small city (Small) and village (Village).

Whereas Table 2 provides information on the place of birth of the descendants of the ancestors, it does not describe the origin and destination of migratory flows. To that end, we build a transition matrix for each generation that gives the percentage of descendants that were born in each category of municipalities, conditionally on the birthplace of the ancestors. The transition matrix of each generation is provided in Table 3. There is an overwhelming majority of sedentary among the descendants of people who were born in villages. However, the proportion of sedentary people declines slightly from one generation to the next, from 96% for children to 85% for great-grandchildren. These declines are offset by increases in proportions in medium and large cities. For ancestors born in large and medium-sized cities, the pattern is different. The share of sedentary among the descendants is relatively less important, compared to ancestors born in villages. In the mean time, the percentage of descendants born in villages grows over generations: it goes from 26% for the children and rises to 50% for the great-grandchildren.

It should not be overlooked that the majority of individuals are born in villages for each generation. During the 20th century, the size of cities changed considerably. In particular, small towns in the periphery of larger ones (the distinction between small, medium, etc. is made at the beginning of the century, around 1800). For example, the village of Ivry, in the neighborhood of Paris (see Bastié, 1964) started with 1,008 inhabitants in 1800, 3,959 in 1836, 10,199 in 1866 up to 33,198 in 1906. Technically, over the century, more than 32 thousand people moved to this 'village'. Similarly, Pantin had 926 inhabitants in 1800, 25,586 in 1900, while Saint-Ouen started with 602 people, and ended with 30,715.[14] Some people came from small villages in the countryside, but also some people have been rejected from the city centre (as discussed in Daumas and Payen, 1976 and Bonvalet and Tugault, 1984).

# 6 Migrations From and Towards the Capital

As the previous sections show, most individuals have remained sedentary from one generation to the next. Among people born in a different place from their ancestors, Section 4

[14]Source: http://cassini.ehess.fr/cassini/fr/html/1_navigation.php.

Table 3: Conditional Transitional Movements Between Large, Medium, Small Cities, and Villages, in Percentages.

| | Children | | | | | Grandchildren | | | | | Great-grandchildren | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | M | S | V | Tot. | L | M | S | V | Tot. | L | M | S | V | Tot. |
| L | 67.10 | 4.14 | 2.42 | 26.34 | 100 | 52.67 | 7.47 | 2.93 | 36.93 | 100 | 36.73 | 9.97 | 3.27 | 50.03 | 100 |
| M | 5.75 | 67.97 | 1.96 | 24.32 | 100 | 10.42 | 47.99 | 3.23 | 38.36 | 100 | 14.02 | 33.85 | 3.02 | 49.11 | 100 |
| S | 4.63 | 4.11 | 65.37 | 25.89 | 100 | 10.21 | 7.64 | 37.46 | 44.69 | 100 | 12.50 | 11.13 | 21.38 | 54.99 | 100 |
| V | 1.13 | 1.52 | 1.29 | 96.06 | 100 | 3.62 | 3.92 | 2.39 | 90.07 | 100 | 5.46 | 5.91 | 3.22 | 85.41 | 100 |

Note: The tables contain a transitional matrix for each generation(children, grandchildren, great-grandchildren) providing the frequency of births of descendants in a large (L) city, in a medium (M) city, in a small (S) city or a village (V) (in column), conditional on the birthplace of the ancestors (in rows).

showed that for a significant proportion of them, the distance travelled remains small, within a radius of 20 km. This section pays particular attention to these individuals, focusing on migration at a micro level, focusing on a special case, that of Paris, the capital of France.[15]

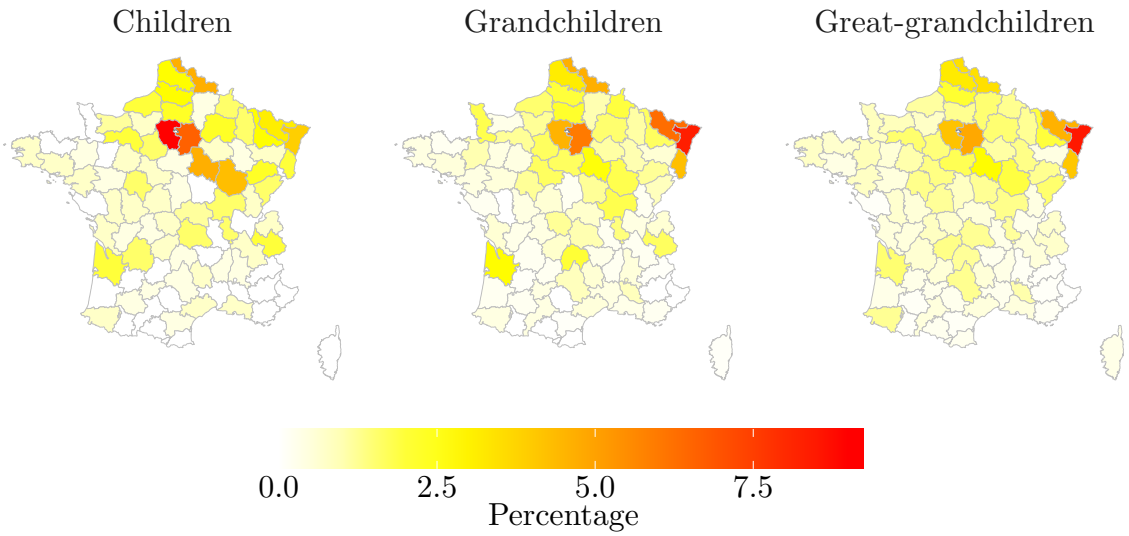
First, we examine the origin of the ancestors of individuals born in Paris, at the departmental level. This is illustrated in Figure 6. For each generation (children on the left, grandchildren in the middle and great-grandchildren on the right), the maps show for each department, each department's share of the ancestors' origin. For clarity purposes, due to the strong sedentariness, the department of Paris has been removed from the calculation. To simplify the understanding, let us take an example. On the left map, the color of the northernmost department (the department called Nord) indicates a value of 4.69%. This reflects the fact that 4.69% of the ancestors of the generation of children born in Paris (*i.e.*, 12 individuals), are from the North (excluding the ancestors born in Paris from the calculation). Three departments in eastern France, namely Moselle, Bas-Rhin and Haut-Rhin, stand out for their relatively high percentage compared to the other departments, for the grandchildren and the great-grandchildren. As shown in Figure 1, a large proportion of the grandchildren in the database were born around the 1970s, when Napoleon III's French Second Empire went to war with the Kingdom of Prussia. In accordance with the Treaty of Frankfurt signed in 1871, Alsace-Lorraine, the territory of which corresponded in part to the three departments mentioned above, was ceded to the German Empire. The inhabitants then had the choice of preserving their French nationality, but to do so, they had to leave the region. These events of political and economic degradation can be put forward to explain the strong representation of ancestors from Moselle, Bas-Rhin and Haut-Rhin among the ancestors of individuals of the generation of grandchildren to be born in Paris. This observation persists in the next generation, as a reinforcing effect.

Secondly, we look at the percentage of descendants born in Paris for each department. This is shown in Figure 7.[16] We note that for individuals of the first generation,

[15]The analysis of migration at the level of the municipalities is in principle made possible by the individual nature of the data. It should be noted, however, that the initial restrictions in this article (intended to focus only on individuals born in the early 19th century and their descendants) only provide a glimpse of migration at the micro level. Indeed, since annual births in small villages are not very numerous, there are few movements observable on such a small scale, with the exception of Paris. Collaborative genealogical data could, however, prove to be an excellent means of observing migrations on a very fine spatial scale, provided that a larger population of ancestors is considered than that chosen in this study.

[16]Again, we exclude the department of Paris for visual concerns, so that the high percentage associated with this department does not inflate the scale.

Figure 6: Department of Birth of Ancestors of Descendants Born in Paris.



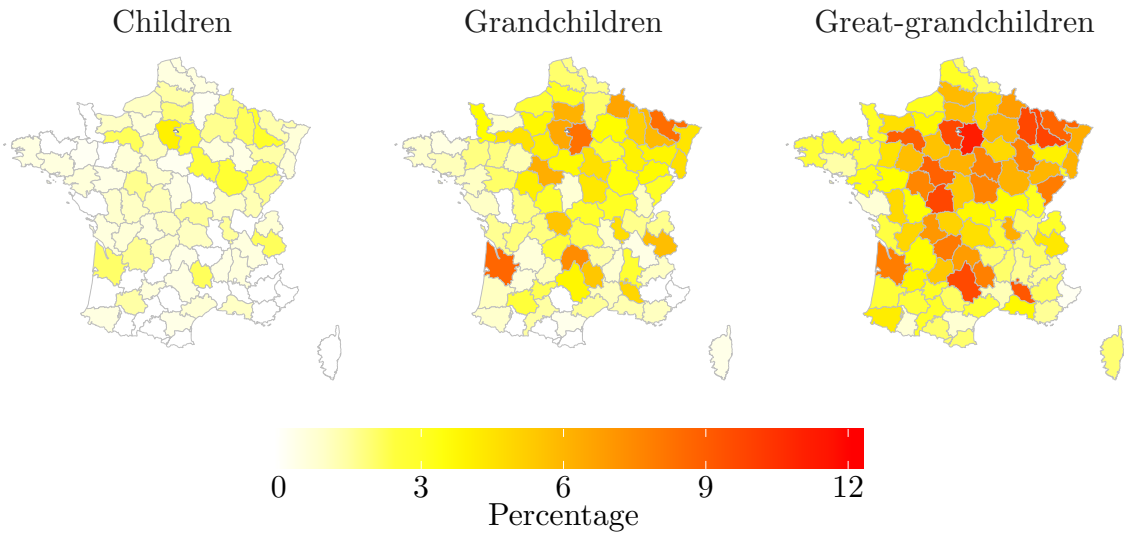

Note: The maps show, for each generation (children, grandchildren and great-grandchildren), the share of each department of birth of ancestors of descendants born in Paris, excluding ancestors born in Paris. The percentage values are represented on a spectrum that ranges from yellow (low values) to red (higher values).

*i.e.*, children, the percentage of those born in Paris is low for most departments. On average, 0.88% of the children of individuals born between 1800 and 1804 were born in Paris per department (excluding the descendants of the ancestors born in Paris). This figure corresponds to a total number of 528 children and an average of 5.87 children by departments. More movements to Paris can be observed in the next generation. The departmental average of descendants in Paris goes up to 2.73% for grandchildren (*i.e.*, 1260 grandchildren, and an average of 14 grandchildren by departments). Some departments stand out for their higher values, notably 'Gironde', where 8.65% of grandchildren (*i.e.*, 27 grandchildren) were born in Paris. We also note that the southern and western departments of France are characterized by low values. Finally, for the generation of great-grandchildren, the departmental average of births in Paris rises to 4.69% (*i.e.*, 3612 great-grandchildren, and an average of 40.1 great-grandchildren by department), with a maximum of 11.2% (*i.e.*, 152 great-grandchildren) for 'Seine-et-Marne'.

These maps seem to have a distance effect: the closer the department is to Paris, the higher the proportion of descendants to be born in the capital. This is in fact what we can observe in Figure 8, for which we plot the percentage of descendants in Paris as a function of the distance from the original departments of the ancestors to the capital.[17] Negative trends support the idea that the closer the departments are to Paris, the higher the proportion of descendants to be born there, whatever the generation. The correlation with the logarithm of the distance is rather small (roughly 25%), but significant. In addition, the more we advance in the generations, the larger the average proportion is. These patterns can be related to popular gravitation models. Ravenstein (1889) introduced a gravity model concept to social sciences, while studying internal migration flows during the 19th century, as discussed in Haynes et al. (1984).

[17]The distance between the regions and Paris is calculated using the distance separating the centroids from the regions.

Figure 7: Percentage of Migrants to Paris.




Note: The maps show the percentage of descendants (of individuals born in metropolitan France between 1800 and 1804) who were born in Paris, for each generation (children, grandchildren and great-grandchildren) and each department (with exception of Paris). The percentage values are represented on a spectrum that ranges from yellow (low values) to red (higher values).

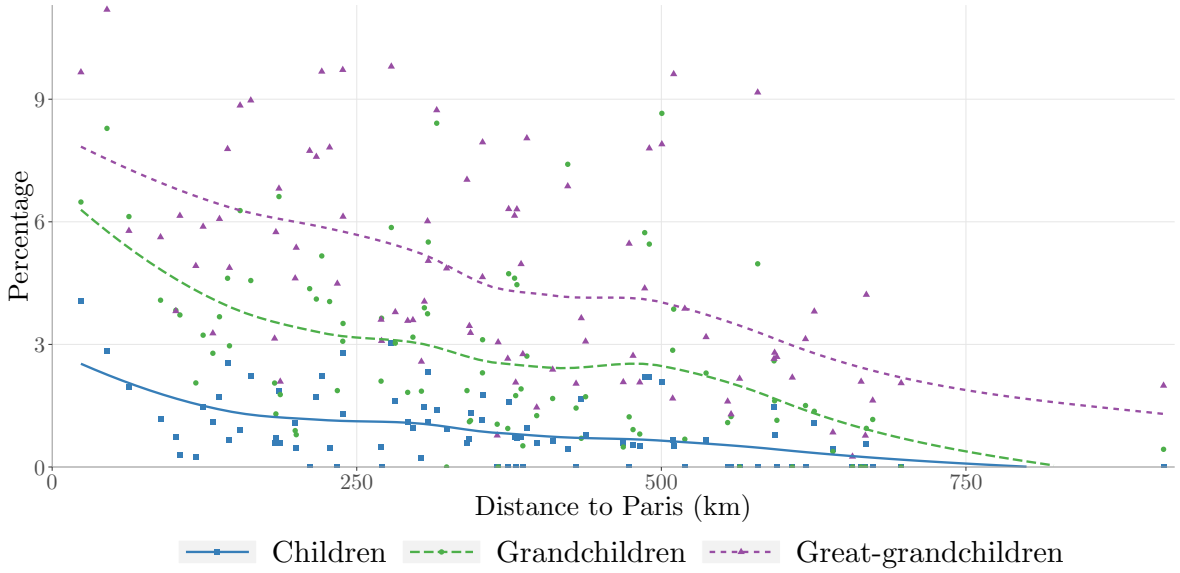
To continue with the idea of the effect of distance on migration in Paris, it is possible to look at what is happening on a smaller spatial scale, now looking at migration out of Paris. We specifically examine transient movements between Paris, its crown and the rest of France. The crown is defined here as the area around Paris within a radius of 20 km from its limits. We report in Table 4 conditional transition matrices for each generation, which indicate, conditionally to the place of birth of the individuals (in Paris or in its Crown), the place of birth of the descendants (in Paris, in its Crown or elsewhere). When the ancestors were born in Paris, the proportion of their descendants also born in Paris is significant, but decreases from generation to generation, going from 79.79% of children to 43.45% of great-grandchildren. While an increasing proportion of the descendants were born in the Paris suburbs, an equally growing, but relatively larger, proportion was born beyond the 20 km surrounding Paris. While one observes progressive movements of departures from the capital, opposite flows are observed in the crown. While only 4.36% of the children of individuals born in the Paris suburbs were born in the capital, this percentage rose in subsequent generations to 11.79% for great-grandchildren.

Table 4: Conditional Transitional Movements Between Paris and its Suburbs.

| | | Children | | | | Grandchildren | | | | Great-grandchildren | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Paris | Suburbs | Other | Tot. | Paris | Suburbs | Other | Tot. | Paris | Suburbs | Other | Tot. |
| Ancestors | Paris | 74.79 | 3.56 | 21.64 | 100 | 55.04 | 9.07 | 35.89 | 100 | 43.45 | 8.98 | 47.58 | 100 |
| | Suburbs | 4.36 | 79.65 | 15.99 | 100 | 6.77 | 68.42 | 24.81 | 100 | 11.79 | 58.19 | 30.02 | 100 |

Note: The table contains a transition matrix for each generation (children, grandchildren, great-grandchildren) that gives the frequency of births in Paris, in its suburbs (<20 km) or further away (in columns), conditional on the birthplace of the ancestors (in lines).

Figure 8: Regional Proportions of Descendants Born in Paris According to the Distance Separating Each department to the Capital.




Note: Each dot represents the share of descendants of a department (blue for children, turquoise for grandchildren, and mauve for great-grandchildren) born in Paris. The distance of a department to Paris (represented on the x-axis) chosen here is that which separates the center of this department from that of the Paris region. The lines correspond to a *loess* smoothing.

# 7 Conclusion

The conceptual analytical framework introduced by Louis Henry in the post-war period, combining the use of registers and the construction of family forms, has contributed to a better knowledge of the population in the past. Representative samples of the population were compiled as an extension of this methodology, such as the TRA survey in France or the Historical Sample of the Netherlands survey. The constitution of such samples requires a considerable effort and takes a lot of time. In this paper, we have used crowd-sourced genealogy data that provide the same type of information as the previously mentioned surveys: dates and places of birth, marriage and death; and, like the TRA survey, family ties between individuals. More specifically, we used family trees from 238,009 users of the Geneanet website. We studied the migration of individuals born in metropolitan France between 1800 and 1804 and that of their descendants over three generations, *i.e.* 141,881 individuals for whom information on place and date of birth was available. Our results are consistent with those of the literature traditionally based on the parish or civil status registers. They suggest that crowd-sourcedt collection of genealogical data, while not as rigorous as that of historians and demographers, can be used to study the population in the past.

# References

Bastié, J. (1964). *La croissance de la banlieue parisienne [The growth of the Paris suburbs]*, volume 17. Presses universitaires de France.

Bean, L. L., May, D. L., and Skolnick, M. (1978). The Mormon historical demography

project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 11(1):45–53. doi:10.1080/01615440.1978.9955216.

Blayo, Y. (1975a). La mortalité en France de 1740 a 1829 [Mortality in France from 1740 to 1829]. *Population (French Edition)*, 30:123. doi:10.2307/1530647.

Blayo, Y. (1975b). Mouvement naturel de la population française de 1740 a 1829 [Natural movement of the French population from 1740 to 1829]. *Population (French Edition)*, 30:15. doi:10.2307/1530644.

Blayo, Y. and Henry, L. (1967). Données démographiques sur la Bretagne et l'Anjou de 1740 à 1829 [Demographic data on Brittany and Anjou from 1740 to 1829]. *Annales de démographie historique*, 1967(1):91–171. doi:10.3406/adh.1967.955.

Bonneuil, N., Bringé, A., and Rosental, P.-A. (2008). Familial components of first migrations after marriage in nineteenth-century France. *Social History*, 33(1):36–59. doi:10.1080/03071020701833325.

Bonvalet, C. and Tugault, Y. (1984). Les racines du dépeuplement de paris [The roots of the depopulation of Paris]. *Population (French Edition)*, 39(3):463. doi:10.2307/1532898.

Bourdieu, J., Kesztenbaum, L., Postel-Vinay, G., and Tovey, J. (2014). The TRA project, a historical matrix. *Population*, 69(2):191–220. doi:10.3917/popu.1402.0217.

Bourdieu, J., Postel-Vinay, G., Rosental, P.-A., and Suwa-Eisenmann, A. (2000). Migrations et transmissions inter-générationnelles dans la France du XIXe et du début du XXe siècle [Intergenerational migration and transmission in 19th and early 20th century France]. *Annales. Histoire, Sciences Sociales*, 55(4):749–789. doi:10.3406/ahess.2000.279879.

Brunet, G. and Vézina, H. (2015). Les approches intergénérationnelles en démographie historique [Intergenerational approaches in historical demography]. *Annales de démographie historique*, 129(1):77. doi:10.3917/adh.129.0077.

Chamoux, A. (1972). La reconstitution des familles : espoirs et réalités [Family rebuilding: Hopes and realities]. *Annales. Économies, Sociétés, Civilisations*, 27(4):1083–1090. doi:10.3406/ahess.1972.422582.

Charpentier, A. and Gallic, E. (2016). Kernel density estimation based on ripley's correction. *GeoInformatica*, 20(1):95–116. doi:10.1007/s10707-015-0232-z.

Cohen, W., Ravikumar, P., and Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78.

Cummins, N. (2017). Lifespans of the European elite, 800–1800. *The Journal of Economic History*, 77(02):406–439. doi:10.1017/s0022050717000468.

Daumas, M. and Payen, J. (1976). *Évolution de la géographie industrielle de Paris et sa proche banlieue au XIXe siècle [Evolution of the industrial geography of Paris and its suburbs in the 19th century]*, volume 1. Conservatoire des arts et métiers: École des hautes études en sciences sociales.

Dribe, M. (2003). Migration of rural families in 19th century southern sweden. a longitudinal analysis of local migration patterns. *The History of the Family*, 8(2):247–265.

Fauve-Chamoux, A. (2004). *Domestic service and the formation of European identity: understanding the globalization of domestic work, 16th-21st centuries*. Peter Lang.

Fine, A. (1991). *La population française au XIXe siècle [The French population in the 19th century]*, volume 1420. Presses Universitaires de France-PUF.

Fire, M. and Elovici, Y. (2015). Data mining of online genealogy datasets for revealing lifespan patterns in human population. *ACM Transactions on Intelligent Systems and Technology*, 6(2):1–22. doi:10.1145/2700464.

Fleury, M. and Henry, L. (1958). Pour connaitre la population de la France depuis Louis XIV. Plan de travaux par sondage [To know the population of France since Louis XIV. Planning of work by sampling.]. *Population*, 13(4):663. doi:10.2307/1525088.

Fourastié, J. (1986). Note sur l'histoire démographique de douelle (lot) 1676-1914 [Note on the demographic history of Douelle (Lot) 1676-1914]. *Population (French Edition)*, pages 483–496. doi:10.2307/1532804.

Gavrilov, L. A., Gavrilova, N. S., Olshansky, S. J., and Carnes, B. A. (2002). Genealogical data and the biodemography of human longevity. *Social Biology*, 49(3-4):160–173. doi:10.1080/19485565.2002.9989056.

Gavrilova, N. S. and Gavrilov, L. A. (2007). Search for predictors of exceptional human longevity. *North American Actuarial Journal*, 11(1):49–67. doi:10.1080/10920277.2007.10597437.

Greenwood, M. J. (1997). Chapter 12 Internal migration in developed countries. In *Handbook of Population and Family Economics*, pages 647–720. Elsevier. doi:10.1016/s1574-003x(97)80004-9.

Haynes, K. E., Fotheringham, A. S., et al. (1984). *Gravity and spatial interaction models*, volume 2. Sage Beverly Hills, CA.

Heffernan, M. J. (1989). Literacy and geographical mobility in nineteenth century provincial france: some evidence from the département of ille-et-vilaine. *Notes*, 1816(20):1872–76.

Henry, L. (1956). Anciennes familles genevoises. etude démographique: XVIème - XXème siècle [Former Geneva families. Demographic study: 16th - 20th century]. *Population (French Edition)*, 11(2):334. doi:10.2307/1524668.

Henry, L. and Blayo, Y. (1975). La population de la France de 1740 a 1860 [The population of France from 1740 to 1860]. *Population (French Edition)*, 30:71. doi:10.2307/1530646.

Ho., J. (1971). Les migrations intérieures en France à la fin du XVIIIe et au début du XIXe siècle [Internal migration in France at the end of the 18th and beginning of the 19th century]. *Population (French Edition)*, 26(4):743. doi:10.2307/1530646.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Price, A. L., and Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*. doi:10.1126/science.aam9309.

Kesztenbaum, L. (2008). Cooperation and coordination among siblings: Brothers' migration in France, 1870–1940. *The history of the Family*, 13(1):85–104. doi:10.1016/j.hisfam.2008.01.006.

Kesztenbaum, L. (2014). L'étude des migrations grâce aux registres matricules militaires [The study of migration through the identification military record]. *Popolazione e storia*, 14(2):9–38.

Lemercier, C. and Rosental, P.-A. (2000). "Pays" ruraux et découpage de l'espace: les réseaux migratoires dans la région lilloise au milieu du xixe siècle [Rural territories and spatial division: migration networks in the Lille region in the middle of the 19th century]. *Population (French Edition)*, 55(4/5):691–725. doi:10.2307/1534691.

Lindahl-Jacobsen, R., Hanson, H. A., Oksuzyan, A., Mineau, G. P., Christensen, K., and Smith, K. R. (2013). The male–female health-survival paradox and sex differences in cohort life expectancy in Utah, Denmark, and Sweden 1850–1910. *Annals of epidemiology*, 23(4):161–166. doi:10.1016/j.annepidem.2013.02.001.

Lucassen, J. and Lucassen, L. (2009). The mobility transition revisited, 1500–1900: what the case of europe can offer to global history. *Journal of Global History*, 4(03):347. doi:10.1017/s174002280999012x.

Matthijs, K. and Moreels, S. (2010). The antwerp COR*-database: A unique Flemish source for historical-demographic research. *The History of the Family*, 15(1):109–115. doi:10.1016/j.hisfam.2010.01.002.

Oris, M. (2003). The history of migration as a chapter in the history of the european rural family: An overview. *The History of the Family*, 8(2):187–215. doi:10.1016/s1081-602x(03)00026-5.

Pinède, C. (1954). L'émigration des habitants du lot en amérique du sud à la fin du xixe siècle [The emigration of the inhabitants of the Lot to South America at the end of the 19th century]. *Revue géographique des Pyrénées et du Sud-Ouest. Sud-Ouest Européen*, 25(4):277–292. doi:https://doi.org/110.3406/rgpso.1954.1386.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, 48(2):167–235. doi:https://doi.org/10.2307/2979181.

Ravenstein, E. G. (1889). The laws of migration. *Journal of the Royal Statistical Society*, 52(2):241–305. doi:https://doi.org/10.2307/2979333.

Rosental, P.-A. (2004). La migration des femmes (et des hommes) en France au XIXe siècle [Migration of women (and men) in France in the 19th century]. *Annales de démographie historique*, 107(1):107. doi:10.3917/adh.107.0107.

Rosental, P.-A. (2006). Between macro and micro: Theorizing agency in nineteenth-century french migrations. *French Historical Studies*, 29(3):457–481. doi:10.1215/00161071-2006-007.

Rosental, P.-A. and Mandelbaum, J. (2003). The novelty of an old genre: Louis Henry and the founding of historical demography. *Population*, 58(1):97–130. doi:10.3917/pope.301.0097.

Shryock, H. S. and Siegel, J. S. (1976). Internal migration and short-distance mobility. In *The Methods and Materials of Demography*, pages 373–405. Elsevier. doi:10.1016/b978-0-12-641150-8.50025-1.

Statistique Générale de la France (2010). Données sur la démographie, la population et l'enseignement primaire sur la période 1800-1925 [Demographic, population and primary education data for the period 1800-1925]. https://www.insee.fr/fr/statistiques/2591293?sommaire=2591397. Accessed December 18, 2018.

Störmer, C., Gellatly, C., Boele, A., and De Moor, T. (2017). Long-term trends in marriage timing and the impact of migration, the netherlands (1650-1899). *Historical Life Course Studies*, 6:40–68.

United Nations (1970). Methods of measuring internal migration. In *Manuals on methods of estimating population*, number 47. United Nations, New York.

# A Construction of Family-Tree From the Data

## A.1 From Ascending to Descending Genealogy

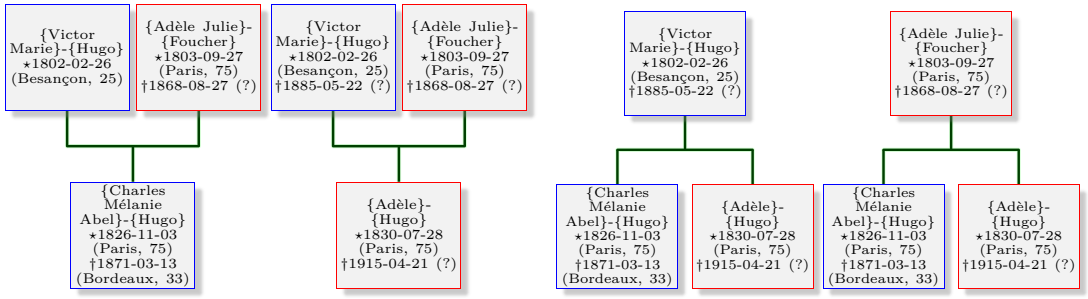
Genealogist mainly uses an ascending strategy to research information on their ancestors (Brunet and Vézina, 2015). They start with an individual and adopt a bottom to the top technique as they look for the ancestors of that individual. The data we get from Geneanet correspond to this ascending system: the record of each individual contains a set of identifiers that allows us to connect him or her to their parents, whenever possible. However, as we want to study the geographical dispersion of the offspring of people who were born between 1800 and 1804, it is wiser to adopt a descending strategy which consists in identifying the descendants of the ancestors. An example is provided in Figure 9a. On the left, we represented two observations from the free of an amateur genealogist. These two observations refer to {Charles Mélanie Abel}-{Hugo} and {Adèle}-{Hugo} who were born in Paris in 1826, and 1830, respectively. Both records point to the same parents: {Victor Marie}-{Hugo} born in Besançon in 1802 and {Adèle Julie}-{Foucher} born in Paris in 1803. Since we want to examine the migration from a generation to the next, we need to retrieve the children of the individuals, not their parents. In our example with the Hugo family, we want to know that Victor Hugo had some children called Charles and Adèle,[18] as illustrated in Figure 9b.

## A.2 Data Cleaning

As each genealogist from the website creates its own family tree, there are a lot of duplicated individuals in the raw data. An ancestor can appear in different user's trees

[18] Actually, Victor Hugo had more children, but they are not represented in Figure 9a for simplicity.

Figure 9: Examples of an Extract of a User's Family-Tree.




(a) Raw Data: Two Individuals Who Share the Same Parents.

(b) What We Want : the Descendants of Each Parent.

with more or less information completed regarding his or her life events. A substantial matching work needs to be done. During this matching stage, we both clean, complete and correct the data. In a nutshell, as the volume of observations is quite large (700 million rows) and as our computer processing capabilities are limited, we split the sample into subsamples, one for each French department.[19] For each department, we identify the duplicated individuals using an algorithm and we reunite them. Once we have applied our algorithm to each department, we regroup the data in a single sample. We screen the data to identify the remaining duplicated individuals and merge them.

Our algorithms that aim at identifying and merge duplicated individuals work in a six-step procedure.

1. We apply a very simple procedure which consists in grouping individuals that share the same following characteristics: last name, first names, gender, date of birth, last name of the mother, last name of the father.

2. We deal with small spelling mistakes in last and first names. We consider that two people who were born in the same department, the same year, who have similar names (e.g., {Ma**tt**hieu Paul}-{Henri} or {Ma**t**hieu Paul}-{Henri}), and whose parents' names are also close, can be identified as the same person. We rely on a string distance measure to evaluate how close two names are to each other. The measure we use is the cosine distance (see Cohen et al. (2003) for more details).

3. We correct gender input errors, using the most frequent declared gender when there is a mismatch between information from different user's family trees.

4. We correct the dates, using the same method as in the third step. Some dates are incomplete in some family trees, where only the year and not the month and day are provided. We complete these dates whenever it is possible, by looking at the information contained in other genealogist's trees.

5. We proceed as in the first step, but this time, instead of looking at the complete set of first names, we only look at the first one that is provided.

[19] Since 1990, French territory is divided in administrative regions called '*départements*'. Currently, there are 101 departments, 96 of which are located in the European territory. We focus on these 96 European French departments.

6. We list the brothers and sisters of any individual from each user's family tree. Among these sororities, we check whether some persons share the same first name and are born or dead the same day. If we find some, we consider that they can be merged as a single person.

# B Calculation of Distances Between Places of Birth

The distances between the birthplaces of an ancestor and his or her descendants are used to study migration between generations. This appendix provides some details on the method used to perform the calculation and provides more information on the composition of the sample.

When the birthplace of an ancestor and that of one of his or her descendants is known, it is possible to use the coordinates of these places to calculate the distance, in kilometers, between the two places. To do this, we use the distm() function of the geosphere package in R. This function uses the Haversine formula to determine the great-circle distance between the two points.

It is important to note that we want to follow the descendants of individuals of people born in metropolitan France between 1800 and 1804, over three generations. Thus, we calculate the distance between the birthplaces of the ancestors and those of the descendants, for each ancestor. When the birthplace of both parents is known, a child will be counted twice in the calculation of distances: once with respect to the mother's birthplace, and once with respect to the father's birthplace. For grandchildren, if we know the birthplace of their four grandparents, this individual will be present four times in the distance data, and so on. Table 5 shows for the descendants of each of the three generations (in columns), the number of individuals for whom we know the place of birth for the number of parents indicated in lines. We read that among children, we know only one of the birthplaces of both parents for 23,135 observations. On the other hand, for 2,762 children, we know the birthplaces of the two ancestors.

Table 5: Number of ancestors for whom the place of birth is known, by generation.

| No. Ancestors | Children | Grandchildren | Great-grandchildren |
|---|---|---|---|
| 1 | 23135 | 27312 | 56254 |
| 2 | 2762 | 4666 | 11180 |
| 3 | - | 198 | 906 |
| 4 | - | 16 | 72 |
| 5 | - | - | 5 |
| 6 | - | - | 0 |
| 7 | - | - | 0 |
| 8 | - | - | 0 |

Note: This table shows the number of known ancestors (in line) for the descendants of each generation (in columns).