

Étude de la démographie française du XIXe siècle à partir de données collaboratives de généalogie

Arthur Charpentier, Ewen Gallic

► **To cite this version:**

Arthur Charpentier, Ewen Gallic. Étude de la démographie française du XIXe siècle à partir de données collaboratives de généalogie. 2018. <hal-01724269>

HAL Id: hal-01724269

<https://hal.archives-ouvertes.fr/hal-01724269>

Submitted on 6 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Étude de la démographie française du XIX^e siècle à partir de données collaboratives de généalogie*

Arthur Charpentier^{a, b} and Ewen Gallic^{†a, c}

^aCREM, UMR CNRS 6211, 7 Place Hoche, 35065 Rennes Cedex, France.

^bUniversité de Rennes 1, 2 rue du Thabor – CS 46510, 35042 Rennes Cedex, France.

^cInstitut Louis Bachelier, Palais Brongniart 28 place de la Bourse 75002 Paris, France.

6 mars 2018

Résumé

Avec l'ère du numérique, les données collaboratives permettent de collecter massivement et à moindre coût des données. Les sites de généalogie fleurissent sur Internet pour proposer à leurs utilisateurs de reconstituer en ligne leur arbre généalogique. Le travail de collecte et de saisie effectué par ces utilisateurs peut potentiellement être réutilisé en démographie historique pour compléter la connaissance du passé de nos ancêtres. Dans notre étude, à partir d'enregistrements concernant 2 457 450 individus français ou d'origine française ayant vécu au XIX^e siècle, nous montrons qu'il est possible de retrouver, bien qu'il subsiste parfois quelques biais, certains résultats de la littérature. Nous proposons d'explorer les caractéristiques temporelles contenues dans les arbres généalogiques pour étudier la longévité, puis de se saisir des caractéristiques spatiales pour analyser les phénomènes migratoires internes en France.

Mots-clés : Généalogie, Données collaboratives, Longévité, Migration.

1 Introduction et motivations

Au milieu du XX^e siècle, les travaux initiés par [Henry \(1956\)](#) fondent le cadre conceptuel d'une nouvelle discipline, la démographie historique. De nombreuses études sont alors

*Ces travaux de recherche ont été réalisés dans le cadre de l'Initiative de Recherche « Valorisation et nouveaux usages actuariels de l'information », placée sous l'égide de la Fondation du Risque en partenariat avec le GENES, l'Université de Rennes 1, l'Université Paris-Est La Vallée. Un travail préliminaire avait été présenté lors des journées « Science XXL » organisées en mars 2017 à l'INED. Nous remercions Olivier Cabrignac et Jérôme Galichon pour leur aide sur l'exploration des données, ainsi que les participants pour les discussions que nous avons pu avoir alors, et qui ont motivé certains éléments présentés dans cette étude.

[†]Auteur correspondant : ewen.gallic@gmail.com

menées dans les années 1960 pour donner un aperçu des caractéristiques de la population française, comme le rappellent [Bourdieu et al. \(2014\)](#). Depuis, la démographie historique a connu un intérêt croissant. Les travaux consacrés à l'enrichissement de la connaissance de nos ancêtres ont fait de la démographie historique un secteur à la croisée de plusieurs disciplines prenant appui sur l'outil statistique : démographie, sociologie, biologie et économie ([Dupâquier, 1981](#)). Les thèmes abordés en démographie portent en particulier sur la reconstitution des effectifs de la population, l'étude de la natalité, de la mortalité ou encore la nuptialité ([Blayo et Henry, 1967](#); [Blayo, 1975a](#); [Henry et Blayo, 1975](#)). En sociologie, il est question de la structure des familles ([Matthijs et Moreels, 2010](#)) et des liens qu'elles peuvent avoir avec la migration sociale ([Bonneuil et al., 2008](#); [van Leeuwen et al., 2015](#); [Bourdieu et al., 2004](#)). La distinction entre ce qui est héréditaire ou lié à l'environnement dans l'explication des caractères et comportements relève plutôt de la biologie, avec en particulier, l'étude des transferts héréditaires liés à l'espérance de vie d'une génération à l'autre ([Gavrilov et Gavrilova, 2001](#); [Motta et al., 2007](#)) et/ou entre frères et sœurs ([Perls et al., 2002](#)). En économie, ce sont les notions de mobilité qui sont principalement étudiées ([Dribe et Lundh, 2005](#)).

Au cœur de la démographie historique se trouvent les données. Leur source varie d'une étude à l'autre. On note une forte exploitation des registres contenant des informations datées sur nos aïeux : dates de naissances, de mariages, de décès. Les registres paroissiaux puis l'état civil depuis la Révolution française, ainsi que les données de recensement s'avèrent largement utilisées, mais ne constituent pas l'unique source d'information présente dans la littérature. Les registres matricules militaires établis par l'administration militaire constituent par exemple une source différente, utilisée notamment pour étudier les migrations au cours de la vie ([Ho., 1971](#); [Kesztenbaum, 2008, 2014](#)). Une source étroitement liée à l'existence des divers registres, puisqu'elle s'appuie sur ceux-là, a participé à l'amélioration de la connaissance de l'histoire de nos ancêtres : il s'agit des données généalogiques. On peut noter que les mormons ont constitué à l'Université d'Utah de volumineuses bases de données généalogiques ayant aidé à l'étude de la population dans le passé ([Bean et al., 1978](#); [Lindahl-Jacobsen et al., 2013](#)). Récemment, avec l'ère du développement des techniques informatiques, ce type de données semble bénéficier d'un gain d'intérêt. La révolution du numérique permet d'accéder dans des temps relativement courts, pour des coûts moindres et des efforts réduits à des quantités conséquentes d'informations relatives à la généalogie. De multiples sites proposent à leurs utilisateurs de reconstituer leurs arbres généalogiques. C'est le cas de [wikitree.com](#), [familysearch.org](#), ou encore [geni.com](#). Pour l'instant, les informations enregistrées par les utilisateurs ont surtout permis d'étudier la longévité des individus ([Gavrilova et Gavrilov, 2007](#); [Gavrilov et al., 2002](#); [Cummins, 2017](#); [Fire et Elovici, 2015](#)).

De manière générale, les données collaboratives semblent offrir de prometteuses perspectives. Elles permettent d'entraîner des modèles, de valider des résultats quasiment en temps réel, et de gérer manuellement des cas dans lesquels des algorithmes ont échoué ([Lease et Yilmaz, 2013](#)). Leur utilisation dans le milieu de la recherche s'observe par exemple en médecine. Dans cette discipline, bien que la collecte des données se restreint parfois à un cercle d'experts, comme dans le cas de la base CIViC¹ visant à accroître les connaissances relatives aux cancers ([Griffith et al., 2017](#)), elle peut aussi être réalisée de manière efficace par des non-experts, pour l'étude du sommeil par exemple ([Warby et al.,](#)

1. Clinical Interpretations of Variants in Cancer : <https://civicdb.org/home>.

2014). En géographie, les données collaboratives, notamment celles d’Open Street Map² (OSM), le projet international visant à créer une carte libre du monde, ont aussi affecté le domaine de la recherche. Leur pertinence était déjà discutée il y a une dizaine d’années par les géographes. En effet, [Haklay \(2010\)](#) ont montré que bien qu’elles comportent quelques erreurs, les informations fournies par les utilisateurs d’OSM sont plutôt précises lorsqu’il s’agit de faire de la cartographie. [Girres et Touya \(2010\)](#) sont arrivés aux mêmes conclusions, en soulignant toutefois des difficultés liées à l’hétérogénéité des données.

La démographie historique peut-elle tirer profit des données de généalogie renseignées par des millions d’utilisateurs ? La réponse à cette question serait positive, à en croire [Kaplanis et al. \(2018\)](#), qui explorent les arbres généalogiques de plusieurs millions d’individus. Leur étude montre que les données de généalogie issues de la collaboration d’amateurs permettent d’obtenir des arbres généalogiques de grande qualité. Nous souhaitons compléter la réponse à l’aide de données de généalogie issues du site [geneanet.org](#), en nous concentrant sur le cas français et en fournissant de plus amples comparaisons avec les résultats déjà connus de la littérature. Nous proposons dans cet article de commencer par un point méthodologique visant à expliquer les méthodes employées pour extraire et classer les informations contenues dans les données de Geneanet ([Section 2](#)). Nous explorons ensuite la représentativité des données ([Section 3](#)), puis examinons deux dimensions de nos données : la dimension temporelle, avec une étude de la mortalité ([Section 4](#)) et la dimension spatiale au travers d’une étude de la migration ([Section 5](#)).

2 Aspects méthodologiques

Ce travail s’appuie sur des données de généalogie issues du site web Geneanet³. Il s’agit de données collectées par des amateurs de généalogie à la recherche de leurs ancêtres. Les utilisateurs du service peuvent construire leur arbre généalogique en renseignant, de manière plus ou moins détaillée, les informations glanées au cours de leurs recherches. Pour chaque individu de l’arbre, les informations concernant trois types d’événements correspondant aux actes d’état civil peuvent être communiqués : (*i*) sa naissance, (*ii*) son ou ses mariages le cas échéant, (*iii*) son décès ; chaque événement pouvant renseigner à la fois le lieu et la date auxquels il est enregistré. Les individus composant un arbre sont reliés en fonction de leurs parents et de leur(s) conjoint(s).

Les données collaboratives de généalogie jouissent de nombreux atouts. La couverture spatiale en est un. De nombreuses études sont contraintes de se limiter à un village ou une région en particulier, faute de données sur un territoire plus vaste ou de moyens financiers pour collecter sur les données sur un ensemble géographique plus large. Ici, les données dont nous disposons permettent de considérer le territoire de la France métropolitaine⁴. Par ailleurs, nous ne sommes pas forcés à réduire notre échantillon à des individus dont le nom de famille commence par un certain n-gramme, comme dans le cas de l’enquête TRA initiée par Jacques Dupâquier en France ou de la base COR en Belgique ([Matthijs et Moreels, 2010](#)). Nous pouvons suivre un nombre plus important d’individus, et ce sur

2. <http://openstreetmap.fr/>

3. <https://www.geneanet.org/>. Les utilisateurs de Geneanet peuvent choisir de publier ou de conserver privé leur arbre généalogique ; les données à notre disposition proviennent uniquement des arbres publiés.

4. En théorie rien n’empêcherait un généalogiste de renseigner un mariage en Belgique ou un décès en Espagne. Bien que nous disposons de certaines observations de descendants nés à l’étranger, nous nous concentrons sur le territoire de la France métropolitaine.

plusieurs générations. Enfin, ce type de données possède un avantage non négligeable en termes de coûts. La numérisation des contenus des divers registres est une tâche à la fois fastidieuse et coûteuse, particulièrement en temps. Le fait de confier ce travail à des centaines de milliers d'utilisateurs confère un avantage non négligeable aux données collaboratives.

Cela dit, les données de généalogie collaboratives comportent quelques limites. L'une d'elles concerne la difficulté à regrouper les branches communes des arbres de différents utilisateurs. Deux freins à ces regroupements peuvent être mis en avant. Le premier est d'ordre pratique ; il résulte de la taille des données. Le volume colossal d'observations nécessite d'avoir une puissance de calcul considérable, ou bien d'avoir recours à des alternatives exigeant de nombreux traitements informatiques. Dans ces derniers, il s'agit de segmenter les données de manière à pouvoir obtenir des objets manipulables tout en faisant attention à ne pas isoler des observations susceptibles d'appartenir *in fine* à la même branche d'un arbre. Le second frein au regroupement est plus délicat. Il concerne la source elle-même des données. La copie des différents registres consultés par les généalogistes peut comporter des erreurs, et la précision des informations retranscrites est inhérente à la volonté de ces individus. Or, le regroupement des branches de deux arbres se fait à l'aide de ces informations recopiées par les internautes. Une erreur de copie peut empêcher d'effectuer le regroupement. Toutefois, certaines erreurs peuvent être corrigées en confrontant les relevés des utilisateurs entre eux, en se fiant aux valeurs les plus fréquemment observées. Par ailleurs, les données de généalogie collaboratives sont loin d'être exhaustives. La présence d'un individu et la quantité d'information à son sujet dans la base sont soumises à l'existence d'une source, ainsi qu'au bon vouloir des usagers. Un lointain arrière grand-oncle qui n'a pas eu d'enfants peut être omis par un généalogiste, faute de temps et d'information précise, par exemple.

La suite de la section propose de décrire la méthode retenue pour la construction des arbres généalogiques, puis examine les biais potentiels contenus dans les données en confrontant l'échantillon de notre analyse avec les résultats de la littérature.

2.1 Construction des arbres généalogiques

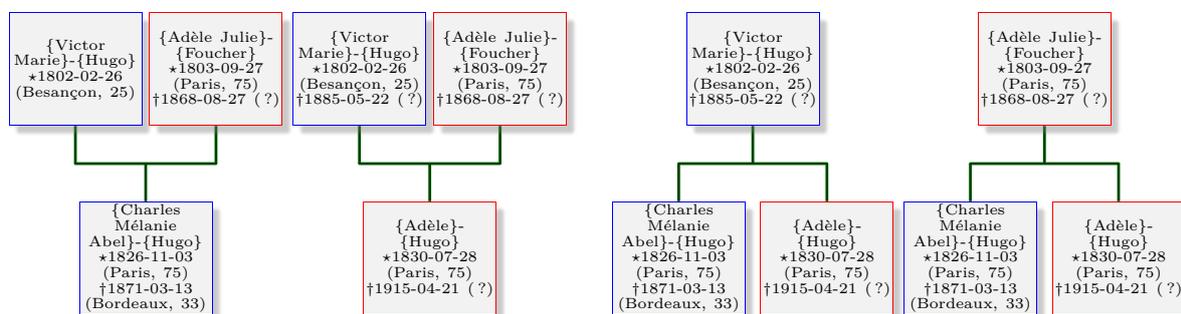
Dans le cadre de cette étude, l'intérêt porte sur les données françaises, qui représentent, selon Geneanet, environ 40% des contenus renseignés par leurs usagers. Nous nous limitons aux individus nés en France entre 1800 et 1804 et à leurs parents et descendants jusqu'à trois générations⁵, soit 2 457 450 personnes. Cette période initiale correspond, comme le rappellent [Fleury et Henry \(1958\)](#), à la reprise sans interruption des relevés de mariages, naissance et décès dans l'ensemble de la France.

Typiquement, les généalogistes amateurs pratiquent principalement la généalogie ascendante pour construire une généalogie individuelle ([Brunet et Vézina, 2015](#)). C'est-à-dire qu'ils partent d'un individu et recherchent progressivement les aïeux. Nos données correspondent à ce type de généalogie : nous disposons d'informations directes concernant les parents – lorsque celles-ci sont disponibles – pour chaque individu présent dans

5. À l'origine, nous disposons d'une liste de 238 009 noms d'utilisateurs ayant mentionné un ancêtre né en France entre 1890 et 1900 dans leur arbre. Nous récupérons les informations de ces arbres, qui contiennent 701 466 921 enregistrements. Nous nous limitons ensuite aux individus nés entre 1800 et 1804, et à leurs descendants (1 547 086 aïeux, 402 190 enfants, 286 071 petits-enfants et 222 103 arrière-petits-enfants).

la base. Or, lorsqu'il s'agit d'étudier la dispersion géographique à partir d'une poignée de femmes et d'hommes, il est plus judicieux d'adopter une démarche différente, consistant à identifier les descendants des individus. Dans ce cas, on parle de généalogie descendante. Il est alors nécessaire de manipuler les données pour suivre les descendants et non les ascendants. Pour clarifier la situation, prenons un exemple concret. Nous avons dans l'arbre généalogique d'un utilisateur deux observations distinctes concernant deux individus : la première concerne {Charles Mélanie Abel}-{Hugo}, né à Paris en 1826 ; la seconde concerne {Adèle}-{Hugo}, née à Paris en 1830. Ces deux observations pointent vers les mêmes parents : {Victor Marie}-{Hugo} né à Besançon en 1802 et {Adèle Julie}-{Foucher} née à Paris en 1803. La Figure 1a illustre ce cas. Ce que nous souhaitons *in fine*, est de pouvoir suivre la descendance de Victor Hugo et d'Adèle Foucher, comme illustré sur la Figure 1b.⁶

FIGURE 1. Exemple d'un extrait d'arbre généalogique d'un utilisateur.



(a) Données type de la base : deux individus (b) Ce que l'on souhaite : les descendants de avec les mêmes parents.

Par ailleurs, un important travail de couplage s'avère nécessaire. En effet, comme chaque amateur de généalogie construit son propre arbre, une même personne peut apparaître dans plusieurs arbres. Il y a *de facto* de nombreux doublons dans les données brutes. Aussi, avant de pouvoir nous intéresser à la description et l'analyse du contenu de notre échantillon comme nous allons le faire dans les section suivantes, nous devons procéder à d'importants nettoyages et mises en formes des données. Avant de décrire ces étapes dans les détails, voici un résumé de notre démarche⁷. Pour commencer, comme le volume d'observations est conséquent et que les capacités de traitement informatique sont limitées, nous découpons l'ensemble de l'échantillon en sous-échantillons par départements français. Puis, pour chaque département, nous repérons les doublons de manière algorithmique. Une fois que chaque département a été traité, nous regroupons les sous-échantillons en un seul, relatif à la France.

Les duplications d'individus s'avèrent utiles pour consolider la qualité des informations. En effet, si un enregistrement est incomplet dans l'arbre d'un utilisateur, les informations manquantes peuvent potentiellement se trouver dans l'arbre d'un autre utilisateur. Aussi, la fusion des doublons permet d'aboutir à une base de données plus riche. Il convient alors de donner quelques détails concernant cette étape. Pour repérer les individus présents dans plusieurs arbres, nous appliquons de manière successive différentes

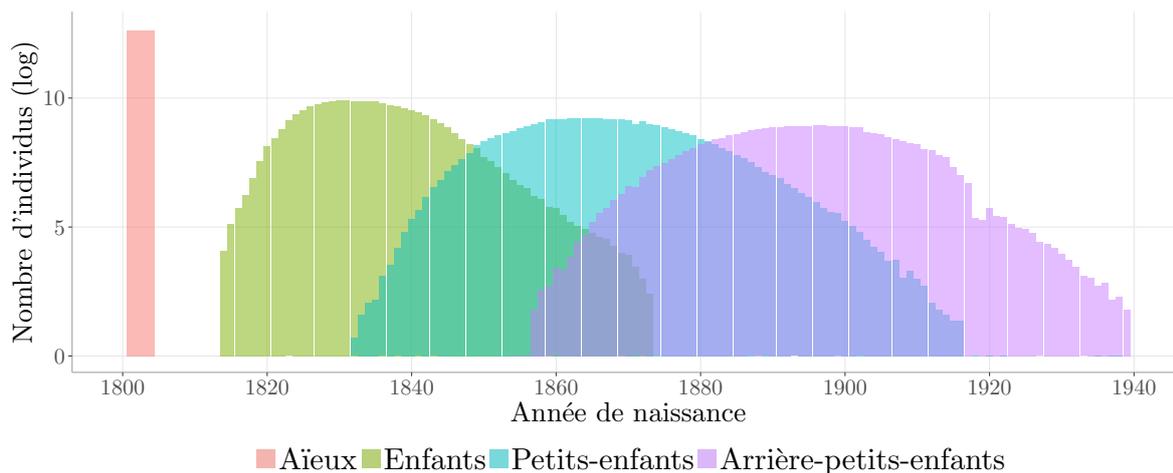
6. On note ici qu'il s'agit d'un extrait de l'arbre d'un utilisateur. Les frères et sœurs de Charles et Adèle Hugo ne sont pas présents sur la Figure 1 par souci de simplicité.

7. De plus amples informations sur la procédure peuvent être obtenues dans l'annexe en ligne : https://3wen.github.io/genealogie_fr/

méthodes. La première est la plus simple. Elle consiste à regrouper les individus possédant les mêmes caractéristiques suivantes : le nom de famille, le prénom, le sexe, la date de naissance, le nom de la mère, celui du père. La seconde méthode de tentative de rapprochement des individus consiste à tenir compte des erreurs de saisie des noms et prénoms. Nous voulons pouvoir considérer que deux personnes nées dans le même département, la même année ayant des noms et prénoms vraiment très proches ($\{\text{Matthieu Paul}\}$ - $\{\text{Henri}\}$ ou $\{\text{Mathieu Paul}\}$ - $\{\text{Henri}\}$ par exemple), ainsi que des noms de parents vraiment très proches puissent être regroupées en une seule. Nous calculons donc des mesures de distance entre les noms et prénoms⁸. Les troisième et quatrième manières de repérer des rapprochements à faire entre les individus consistent respectivement à porter l'intérêt sur les erreurs de sexe et les dates incomplètes (jours ou mois manquants ou erronés). La cinquième façon d'identifier des duplications consiste à s'appuyer sur les rapprochements de prénoms, en ne conservant cette fois que le premier prénom, et non plus les deux ou trois autres. La sixième et dernière manière de repérer les *duplicata* s'appuie sur les frères et sœurs. Nous listons les frères et sœurs de chaque individu, s'il en a. Le cas échéant, nous regardons si parmi eux, certains portent le même prénom et sont nés ou décédés la même année. Si tel est le cas, nous considérons qu'il est possible de les regrouper en une seule et même personne.

Une fois les données des arbres des différents utilisateurs appariées, nous classons chaque individu en fonction de sa génération, en définissant 4 catégories : (i) les aïeux ; (ii) leurs enfants ; (iii) leurs petits-enfants, (iv) leurs arrière-petits-enfants. La distribution de l'année de naissance des individus de chaque catégorie est représentée sur la Figure 2. L'échelle des ordonnées est logarithmique, le nombre d'aïeux étant relativement important comparativement à leurs descendants. Les aïeux sont arbitrairement choisis en retenant les individus nés en France entre 1800 et 1804. Puis, chaque génération naît approximativement avec un intervalle de 30 ans.

FIGURE 2. Distribution des années de naissance de l'échantillon par génération.



8. Plus précisément, nous calculons la mesure cosinus. Pour plus de détails sur cette mesure de distance entre chaînes de caractères, le lecteur est prié de se reporter à l'étude de [Cohen et al. \(2003\)](#).

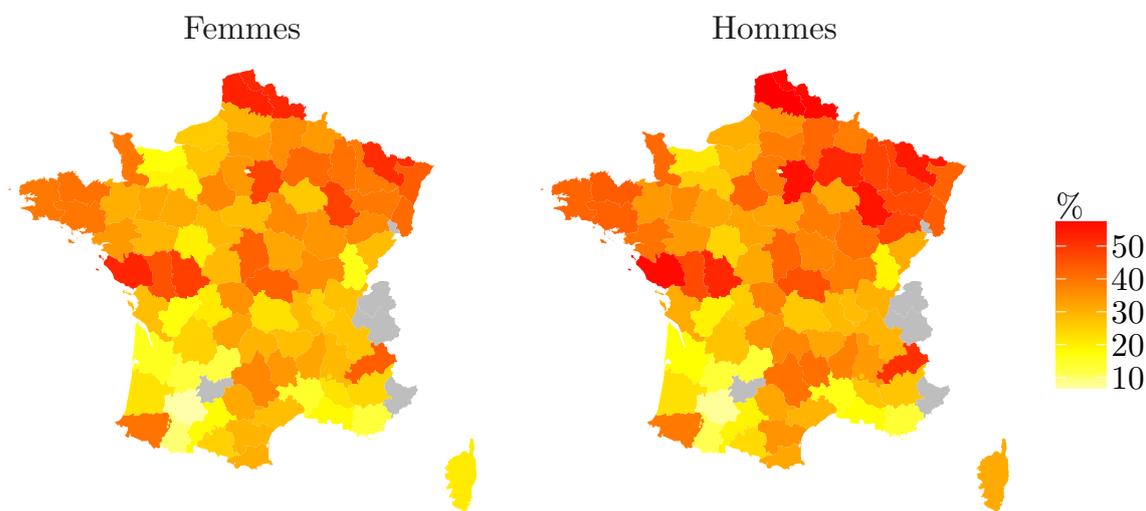
3 Représentativité des données

Il convient de rester prudent face aux résultats proposés par les analyses menées sur notre échantillon. Pour avoir une meilleure idée de l'intérêt que peuvent représenter les données collaboratives de généalogie dans l'étude de la population, nous décrivons notre échantillon et le comparons avec des résultats présents dans la littérature.

3.1 Nombre de naissances

L'échantillon est construit en partant d'individus nés en France entre 1800 et 1804, et en suivant leurs descendants jusqu'à trois générations. Une première mesure de la représentativité des données consiste à comparer le nombre d'individus à l'origine de l'échantillon avec des statistiques institutionnelles. L'INSEE fournit des informations relatives au nombre de naissances par département à différents points dans le temps tout au long du XIX^e siècle ([Statistique Générale de la France, 2010](#)). En 1801, les valeurs sont données en fonction du sexe des nouveaux-nés. Aussi, nous comparons les valeurs de notre échantillon avec celles de l'INSEE, en calculant le rapport entre les deux. Les cartes de la [Figure 3](#) proposent une représentation graphique de ces rapports, à gauche pour les femmes et à droite pour les hommes. Une hétérogénéité dans la représentativité s'observe entre les régions, mais semble indépendante du genre (la corrélation entre la représentativité départementale des hommes avec celle des femmes est de 98%).

FIGURE 3. Proportion des naissances par département dans l'échantillon comparative-ment aux données communiquées par l'INSEE.



Note : Ces cartes montrent la représentativité de l'échantillon par département en comparant le nombre d'individus nés en 1801 enregistrés dans la base aux enregistrements de l'INSEE. La couleur grise indique une valeur manquante.

3.2 Rapport femmes-hommes

L'échantillon final comporte une légère surreprésentation des hommes. Le taux de masculinité à la naissance, c'est-à-dire le rapport entre le nombre de naissances de gar-

çons sur le nombre de naissances de filles, s'établit à 117 pour les individus nés entre 1800 et 1804 dans notre échantillon. Les données de l'INSEE concernant les naissances par sexe indiquent, pour 1801, un taux de masculinité de 105. [Blayo et Henry \(1967\)](#) obtiennent une valeur similaire, *i.e.*, 105,4 pour la période 1740-1829 pour la Bretagne et l'Anjou. Avec l'ensemble de l'échantillon, nous obtenons une valeur moyenne annuelle de 116 garçons pour 100 filles. Cette surreprésentation masculine s'observe à travers l'ensemble des départements, avec un minimum de 106,34 (soit 51.5% d'hommes) dans la Loire et un maximum de 126.84 (soit 55.9% d'hommes) dans les Vosges. L'existence d'un biais sexiste dans les études de généalogie n'est pas nouvelle. [Gavrilov et Gavrilova \(2001\)](#) mentionnent l'existence d'une sous-déclaration des femmes et des enfants dans les bases de données généalogiques en général. Ils estiment que le biais de représentativité féminine qu'ils observent dans leurs propres données des familles royales et nobles d'Europe concerne principalement les enregistrements plus anciens, c'est-à-dire du XIX^e siècle, et moins les plus récentes qui sont plus complètes. L'enquête TRA souffre aussi d'un biais sexiste, comme le mentionnent [Bourdieu et al. \(2014\)](#), bien que dans ce cas précis, il serait en partie dû au mode de constitution de l'échantillon consistant à suivre des branches masculines, en rajoutant toutefois des épouses à chaque génération.

3.3 Taux de fécondité

Le nombre d'enfants par femme dans l'échantillon est inférieur à ce que l'on peut trouver dans la littérature. En effet, les femmes ayant été mariées au cours de leur vie et ayant atteint un âge d'au moins 15 ans ont eu en moyenne 1,46 enfant, d'après les statistiques révélées par l'échantillon. Cette statistique est bien inférieure aux valeurs reportées par [Chesnais \(1986, p. 311\)](#) qui atteignent 4.46 enfants par femme au début du XIX^e siècle et diminuent progressivement jusqu'à 2.9 enfants par femme cent ans plus tard. La faiblesse de cet indicateur peut s'expliquer par le fait que dans les généalogies constituées par des amateurs, comme souligné par [Brunet et Bideau \(2001\)](#), les branches de certains individus, notamment ceux restés célibataires, ne sont pas nécessairement intégrées. L'absence de ces individus impliquerait *de facto* une mesure biaisée du taux de fécondité. Un autre facteur explicatif provient de la manière de composer l'échantillon. Le nombre d'enfants moyen par femme est tiré vers le bas par les individus de la troisième génération, pour qui le comptage du nombre de descendants n'est que rarement possible, du fait de la troncature de l'arbre au-delà de la troisième génération. Pour les femmes de la génération initiale (aïeules), le nombre d'enfants moyen s'établit à 1.36, il grimpe à 1.61 pour celles de la génération suivante (enfants), puis à 1.62 pour les suivantes (petits-enfants).

La distribution du nombre d'enfants par femme mariée en fonction de chaque génération permet un aperçu plus riche. Elle est donnée dans le [Tableau 1](#). Chaque ligne décrit pour une génération le pourcentage de femmes ayant eu le nombre d'enfants indiqué en colonne. On y lit que la part de femmes n'ayant pas eu d'enfant est élevée : 46% pour les individus de la première génération, environ 25% pour les deux générations suivantes. Si la part de femmes sans enfant pour les individus de la génération initiale semble élevée au regard des proportions avancées par [Houdaille et Tugault \(1987\)](#) qui avoisinent plutôt les 15%, celles des générations suivantes coïncident avec les valeurs du tout début du XX^e siècle rapportées par [Toulemon \(1995\)](#).

Tableau 1. Répartition du nombre d'enfants par femme mariée par génération (en pourcentage).

Génération	Nombre d'enfants											
	0	1	2	3	4	5	6	7	8	9	10	>10
0	46,98	24,54	10,26	6,14	4,02	2,62	1,89	1,20	0,85	0,53	0,36	0,63
1	27,36	39,18	13,18	7,48	4,80	2,86	1,87	1,22	0,84	0,51	0,27	0,43
2	25,81	40,33	13,72	7,88	4,50	2,68	1,89	1,17	0,78	0,47	0,30	0,46

4 Dimension temporelle des données : mortalité

Les données de généalogie offrent des informations à première vue simples, qui permettent néanmoins d'explorer de nombreuses pistes concernant la vie de nos ancêtres : les dates associées aux événements consignés dans les registres. En combinant la date de naissance et celle de décès, une étude de la mortalité peut être réalisée. Cette section propose d'explorer les questions de mortalité et de survie pour aïeux, à l'échelle nationale puis régionale.

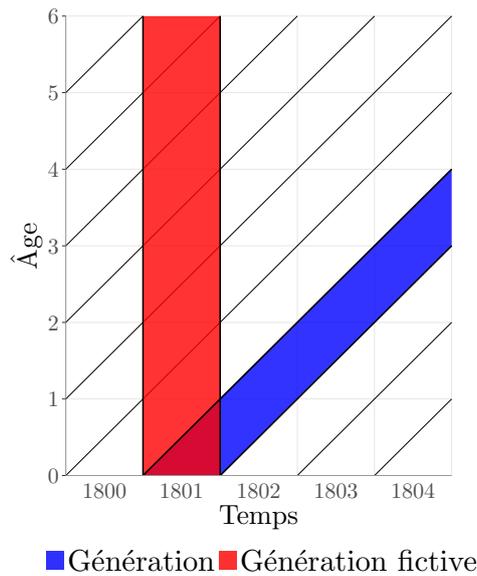
Une particularité de la nature des données est qu'il est possible de suivre des cohortes d'individus pour estimer les probabilités de survie ainsi que l'espérance de vie à un âge donné. Traditionnellement, l'espérance de vie à un âge donné, par exemple à l'âge 0, soit à la naissance, pour une année donnée, est estimée en effectuant une photographie de l'état du monde l'année correspondante, et en relevant les conditions de mortalité à chaque âge de l'année. Graphiquement, cela correspond à calculer l'espérance de vie à partir des valeurs correspondant à la bande verticale rouge représentée sur le diagramme de Lexis proposé à la [Figure 4](#)⁹. Cette méthode de calcul, bien que largement utilisée, ne prend pas en compte le vieillissement des individus d'une cohorte donnée. Nous décidons d'employer un modèle davantage prospectif, consistant à suivre les individus par cohorte, plutôt que de fixer une date puis de relever les conditions de mortalité à cette date. Sur le diagramme de Lexis, cela correspond à lire les informations correspondantes à la bande oblique bleue.

Les cohortes que nous décidons de suivre sont au nombre de 5. Il s'agit de celles des aïeux de la base, soit de 1800 à 1804, incluse. Pour chaque cohorte, en utilisant les dates de naissance et de décès, qui sont indiquées pour 813 551 des aïeux (soit 53% d'entre eux), nous calculons le nombre d'individus encore en vie ainsi que le nombre de décès pour chaque âge. Il est alors aisé de calculer la probabilité de survie à chaque âge, pour chaque cohorte. Nous confrontons nos estimations avec les valeurs issues des tables de mortalité proposées par [Vallin *et al.* \(2001\)](#). Une légère complication apparaît alors, puisque les tables de mortalité de [Vallin *et al.* \(2001\)](#) commencent à partir de 1806, ce qui nous empêche *de facto* de comparer les probabilités de survie sur les premières années de vie des aïeux de notre base¹⁰. Pour contourner ce manque d'informations historiques, nous décidons de restreindre les estimations aux individus ayant survécu au moins jusqu'à leur sixième année. Afin de faciliter la comparaison entre nos estimations et celles de référence,

9. Selon [Vandeschrick \(1992\)](#), la représentation graphique proposée à la [Figure 4](#) serait attribuable à Roland Pressat.

10. Par la suite, nous calculerons des espérances de vie à la naissance à partir des tables de mortalité complètes.

FIGURE 4. Diagramme de Lexis.



Note : Les lignes verticales indiquent les dates, les lignes horizontales délimitent les âges exacts et les lignes obliques croissantes représentent des lignes de vie. La bande oblique bleue représente une génération telle que nous la suivons dans les données tandis que la bande verticale rouge correspond à une génération fictive pour l'INSEE lors du calcul de l'espérance de vie à la naissance.

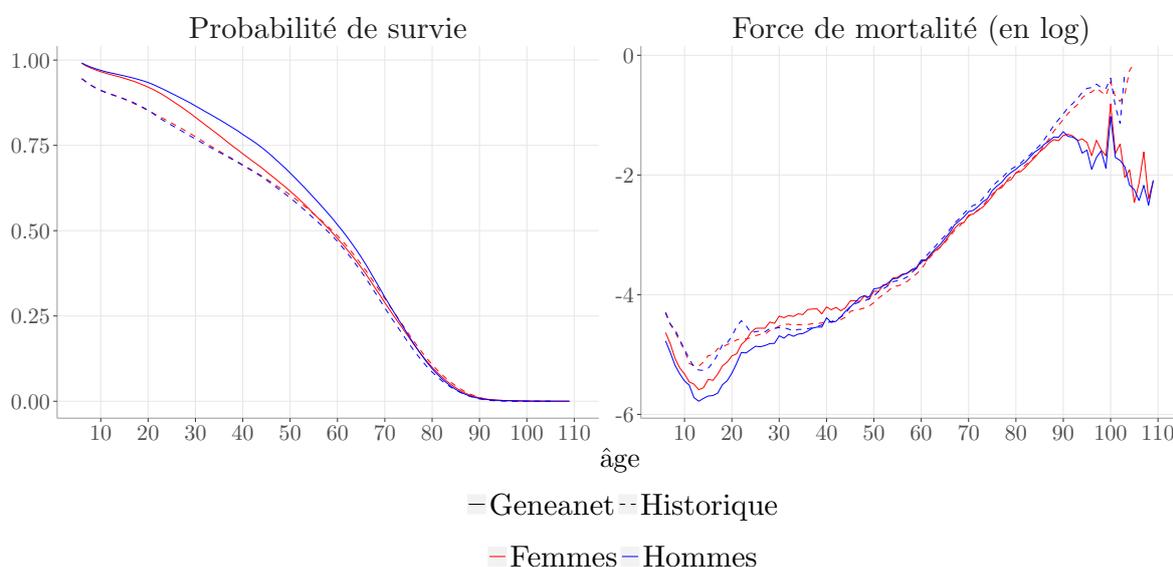
nous effectuons une moyenne par âge des probabilités de survie.

Nos estimations réalisées à partir des données de Geneanet ont tendance à fournir une surestimation de la survie des individus, particulièrement en ce qui concerne les hommes, comme le montre le graphique de gauche de la [Figure 5](#). L'estimation de la force de mortalité, c'est-à-dire de la probabilité de décès instantané conditionnelle à un âge atteint, permet d'apporter un éclairage sur la surestimation de la survie des individus dans nos données. L'on observe en effet sur le graphique de droite de la [Figure 5](#) que nous sous-estimons principalement la probabilité de décès des individus jeunes ainsi que celle des individus très âgés. Pour les individus dans l'entre-deux, *i.e.*, ceux qui ont survécu au moins jusqu'à leur vingt-cinquième anniversaire, mais pas au-delà de leur quatre-vingt-dixième (ce qui représente deux tiers des aïeux) l'estimation est bien meilleure.

En effectuant cette fois l'analyse sur l'ensemble des individus de la base, non plus en nous limitant aux individus ayant survécu au moins jusqu'à leur sixième année, nous pouvons calculer l'espérance de vie à la naissance des individus, pour chacune de nos cinq cohortes. Nous obtenons des valeurs de 41.8 années pour les femmes et 43.5 années pour les hommes, pour les cohortes de 1800 à 1804. Ces espérances de vie à la naissance sont supérieures à celles que l'on peut calculer avec les tables de mortalité de [Vallin et al. \(2001\)](#) pour la cohorte 1806, *i.e.*, la première cohorte complète dans ces tables. En effet, les valeurs obtenues sont de 38.1 années de vie espérée à la naissance pour les femmes et 36.3 années pour les hommes.

Les données de généalogie collaboratives souffrent donc d'un biais dans l'estimation de l'espérance de vie à la naissance. Nous souhaitons savoir si, en revanche, les différences régionales sont plus fidèlement retranscrites. Pour ce faire, nous estimons l'espérance de vie à la naissance dans chacun des départements, de nouveau pour les cohortes d'individus de 1800 à 1804. Un excellent point de comparaison est fourni par [van de Walle \(1973\)](#), qui estime, pour dix regroupements de cohortes différentes de femmes, les espérances de vie

FIGURE 5. Comparaison des fonctions de survie (gauche) de force de mortalité (droite) estimées pour les femmes et les hommes avec les estimations historiques réalisées à partir de tables de mortalité.



Note : Les valeurs historiques sont calculées à partir des tables de mortalité de Vallin *et al.* (2001), disponibles en ligne à l'adresse suivante : <http://www.lifetable.de/cgi-bin/country.php?code=fra>. Les estimations sont réalisées conditionnellement au fait que les individus aient survécu au moins jusqu'à leur sixième année d'existence.

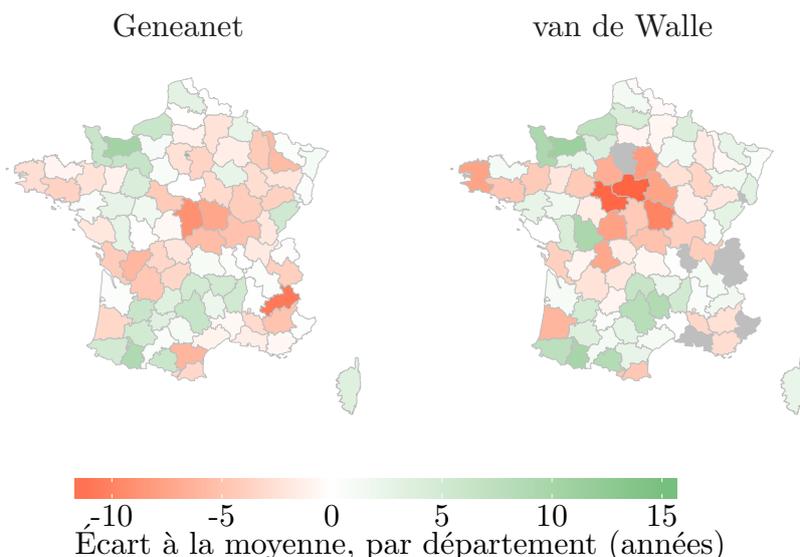
à la naissance par département français. Un de ces regroupements concerne les cohortes de 1801 à 1810, sur lequel nous nous appuyons. Comme nos propres estimations souffrent d'un biais, la comparaison avec celles de la littérature n'est pas directement possible. Nous décidons alors de calculer, pour nos données dans un premier temps, puis pour les données de van de Walle (1973) dans un second temps, les écarts d'espérance de vie par rapport à la moyenne régionale. Les résultats sont proposés sur les cartes de la Figure 6. Globalement, les contrastes régionaux sont plutôt similaires entre les deux sources. On peut toutefois noter une différence au niveau du triangle désigné par van de Walle (1973), dont les pointes sont l'Ille-et-Vilaine, la Nièvre et la Girond, délimitant une région ayant été sujette à une crise de mortalité.

Les résultats de cette dernière comparaison laissent penser qu'il peut être avisé de s'intéresser davantage à l'aspect spatial qu'offrent les données de généalogie collaborative. Aussi, la section suivante s'oriente vers des analyses exploitant les indications géographiques fournies dans les enregistrements de Geneanet.

5 Dimension spatiale des données : sédentarité et migration intérieure en France

Les informations proposées par les données de généalogie permettent non seulement de s'intéresser aux aspects temporels de l'étude de la population passée, mais également de se pencher sur un second aspect tout aussi riche : celui de la spatialité. En effet, outre les dates associées aux événements de naissance, mariage ou de décès, les registres

FIGURE 6. Espérances de vie à la naissance pour les femmes, par département.



Note : Les cartes indiquent les écarts d'espérance de vie à la naissance à la moyenne, par département, pour notre échantillon (gauche) et les estimations de [van de Walle \(1973\)](#) (droite). Les espérances de vie à la naissance sont estimées pour les cohortes de 1800 à 1804 pour nos données, tandis que celles de [van de Walle \(1973\)](#) concernent les cohortes de 1801 à 1810.

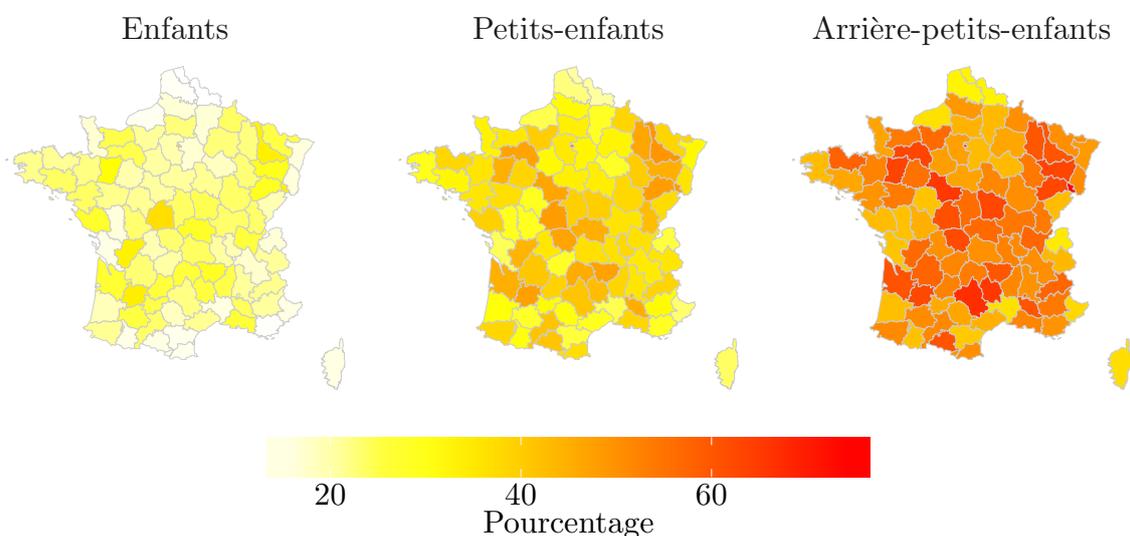
fournissent une indication du lieu, sous forme de coordonnées GPS ¹¹. Une étude de la migration est dès lors envisageable. La migration des individus dans le passé est un des sujets au cœur d'un vaste champ de la littérature en démographie historique. Nous parlons ici de migration interne, laissant les migrations vers l'international de côté. L'on distingue, d'après [Fine \(1991, pp. 88–89\)](#), quatre types de phénomènes migratoires internes que nous pouvons classer en fonction de leur rapport au temps, allant de la migration de courte durée à celle de long terme : *(i)* les migrations alternantes, *(ii)* les saisonnières, *(iii)* celles temporaires et enfin *(iv)* les migrations définitives de glissement. Les migrations alternantes sont liées au marché du travail. Elles caractérisent les déplacements fréquents entre les lieux de domicile et de travail, qui ont lieu généralement quotidiennement. Les migrations saisonnières concernent les déplacements de population en relation avec l'activité économique saisonnière, comme celle des récoltes agricoles en été. Dans ce cas, les individus émigrent de manière momentanée pour loger plus près des endroits dans lesquels l'activité économique s'est temporairement accrue le temps de la saison. Lorsque la saison s'achève, l'offre de travail diminue à nouveau et les travailleurs saisonniers retournent à leur point de départ. Les migrations temporaires concernent des déplacements momentanés au cours d'une vie en lien avec un changement d'activité. Enfin, les migrations définitives de glissement caractérisent les déplacements pérennes d'individus : les montagnards vont remplacer les gens des plaines partis en villes. Les données collaboratives de généalogie ne permettent *a priori* pas d'étudier les deux premières catégories de phénomènes migratoires. Elles pourraient potentiellement s'attacher à décrire les migrations temporaires, en suivant les déplacements entre le lieu de naissance des individus et les lieux dans lesquels ils se marient, ont des enfants, puis décèdent. Malheureusement, pour l'heure, les informations manquantes sont trop nombreuses pour pouvoir suivre de tels parcours de vie. En revanche, à partir des données de naissance des individus, nous

11. Les lieux sont géolocalisés par Geneanet ; ces derniers fournissent des couples de coordonnées (longitude, latitude) pour chaque enregistrement, lorsqu'il a été possible d'identifier le lieu.

pouvons utiliser les données de généalogie afin d'étudier le quatrième type de phénomène migratoire, en suivant de génération en génération où naissent les individus.

Une première façon de décrire les migrations intergénérationnelles consiste à regarder à l'échelle des départements s'il existe des déplacements conséquents de la population. Pour ce faire, nous comparons le département de naissance de nos individus nés entre 1800 et 1804 en France avec celui de leurs descendants. La carte de gauche de la [Figure 7](#) indique que la majorité des enfants des individus à l'origine de l'étude sont nés dans le même département que leur aïeul. La proportion de petits-enfants à naître dans un département différent de celui leur ancêtre augmente légèrement, et il faut attendre la génération des arrière-petits-enfants pour observer des distinctions régionales marquées.

FIGURE 7. Pourcentage de descendants nés dans un département différent de celui de leur aïeul, par département.



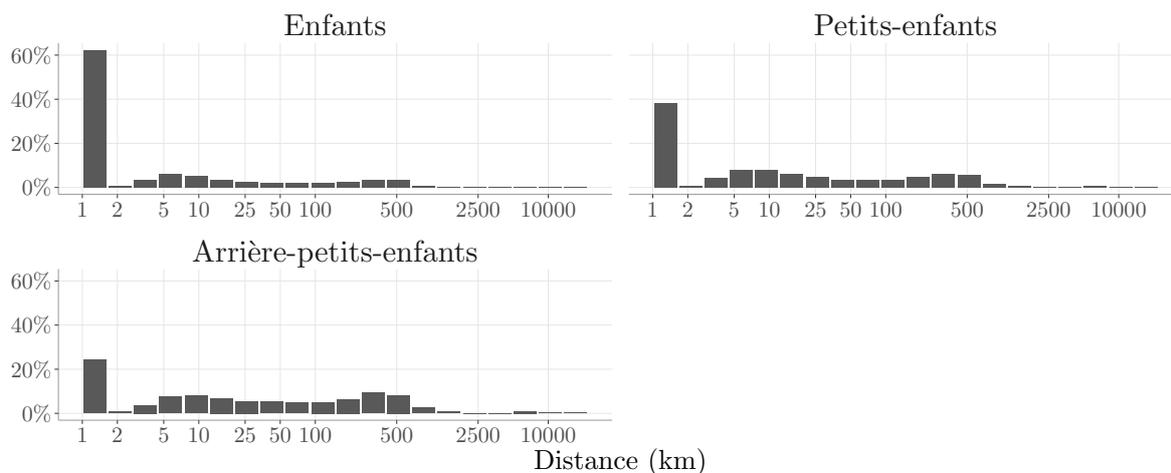
Note : Ces cartes indiquent le pourcentage d'enfants (gauche), petits-enfants (milieu) et arrière-petits-enfants (droite) nés dans un département différent de celui de leur aïeul né entre 1800 et 1804 en France.

5.1 Migrations de courte et longue distance

Une analyse plus fine des déplacements est possible en s'intéressant non plus aux départements de naissance, mais aux distances parcourues par les individus d'une génération à la suivante. Plus précisément, en utilisant les coordonnées des lieux de naissance des individus nés en France entre 1800 et 1804 et en les confrontant à celles des lieux de naissance de leurs descendants, il est aisé de calculer les distances intergénérationnelles parcourues par les français au XIX^e siècle. La [Figure 8](#) propose une visualisation graphique de ces distances. À l'instar des résultats mis en avant par [Bourdieu *et al.* \(2000\)](#), nous observons une bimodalité dans la distribution, avec une forte masse observée pour les sédentaires et une seconde autour d'une dizaine de kilomètres. Nous constatons également une diminution progressive au fil des générations du pourcentage de sédentaires¹².

12. De plus amples informations sur les distances parcourues en fonction du genre des individus sont reportées dans la [Tableau 6](#) en annexe.

FIGURE 8. Migration entre les générations.



Note : Les graphiques montrent la distribution des distances séparant le lieu de naissance des individus de l'échantillon du lieu de naissance de leur plus proche parent, pour chaque génération.

Dans la littérature, il est coutume de caractériser les déplacements des individus en fonction de la distance parcourue, en séparant les courtes des longues distances. Selon Rosental (2004), au-delà d'une certaine distance de migration, les individus quittent un environnement plus ou moins familier, propre à leur commune. L'allongement de cette distance entraîne dans le même temps, comme le fait remarquer Kesztenbaum (2008), un accroissement des coûts, qu'ils soient économiques ou non. Ces coûts deviennent alors un frein à la migration, que certaines catégories de la population sont plus enclines à supporter. Les travaux de la littérature mettent en exergue l'existence d'un phénomène de sélection positive des individus prenant part à la migration de longue distance. Un premier facteur de sélection concerne le lieu de vie des individus, et oppose les milieux urbains aux milieux ruraux. En effet, comme le rappelle Rosental (2004), il existe une différence marquée entre les populations urbaines et rurales au XIX^e siècle, les populations urbaines ayant tendance à être plus attirées par les villes et plus disposées à parcourir de longues distances. L'éducation est un autre facteur mis en avant, notamment par Bourdieu *et al.* (2000) qui montrent que les migrations de proximité sont associées à des niveaux éducatifs relativement modestes comparés aux migrations de longue distance. Bonneuil *et al.* (2008) associent ces dernières à la mobilité sociale.

Pour effectuer la distinction entre les migrations de courte de celles de longue durée, il convient alors de déterminer une distance maximale au-delà de laquelle une migration ne pourra plus être considérée comme courte. Nous retenons une valeur de 20km. À titre de comparaison, Rosental (2004), Bourdieu *et al.* (2000) et Kesztenbaum (2008) choisissent une valeur de 25km, 20km et 17km respectivement. Leur choix, tout comme le nôtre, se fonde *grosso modo* sur la valeur médiane des distances parcourues par les migrants. Comme indiqué dans la Tableau 2, la part de migrants de proximité est quasiment égale à celle de migrants de longue distance (19.43% et 18.40%, respectivement) pour les enfants des individus de la première génération. Par ailleurs, le choix d'une distance de 20 km offre aussi l'avantage de rester dans un voisinage familier pour les migrants de courte distance.

Le Tableau 2 nous montre que la part des enfants des individus nés au début du XIX^e siècle qui naissent dans le même endroit que leur parent s'élève à 62%, qu'elle diminue à

38% pour les petits-enfants et chute à 24% pour les arrière-petits-enfants. Dans le même temps, nous notons que les migrations longues ont quant à elles progressé nettement plus, relativement aux migrations courtes.

Tableau 2. Déplacements intergénérationnels en France, en pourcentages.

	Enfants	Petits-enfants	Arrière-petits-enfants
Sédentaires	62,17	38,06	24,17
Migrants de proximité	19,43	27,70	27,06
Migrant de longue distance	18,40	34,24	48,77

Note : Les sédentaires sont nés dans le même lieu que leur aïeul, les migrants de proximité sont nés à moins de 20 km d'écart et les migrants de longue distance à plus de 20 km.

5.2 Transitions entre les grandes et petites villes

À l'aide des coordonnées géographiques fournies dans notre échantillon, il est aisé de retrouver la commune de naissance des individus. Nous considérons que la commune de naissance renseignée correspond au lieu de vie des parents à ce moment. Le risque d'une sur représentativité par rapport à la réalité des grandes communes est limité pour cette époque, puisque comme l'indique (Fine, 1991, p. 34), une grande majorité des accouchements (98%) s'effectuent à domicile à la veille de la Seconde Guerre mondiale.

Des informations relatives aux communes, comme la taille de leur population, peuvent ensuite venir enrichir l'échantillon. Des tableaux de recensement de la population sont notamment disponibles à l'échelle des chefs-lieux et villes (Statistique Générale de la France, 2010). Nous nous appuyons sur ces données, qui concernent les chefs-lieux et villes pour définir quatre catégories, que l'on retrouve pour des raisons pratiques d'échantillonnage chez Fleury et Henry (1958) ou plus tard Blayo (1975b) :

- grandes villes : plus de 50 000 habitants ;
- villes moyennes : entre 10 000 et 50 000 habitants ;
- petites villes : moins de 10 000 habitants ;
- très petites villes : les villes qui ne sont pas présentes dans les registres de la Statistique Générale de la France.

En utilisant les statistiques de la population de 1836, les grandes villes sont au nombre de 9 : Paris, Lyon, Marseille, Bordeaux, Rouen, Toulouse, Nantes, Lille et Strasbourg. Les villes moyennes sont au nombre de 98 ; les petites sont 254 ; les très petites constituent le reste. Le découpage effectué nous donne une majorité de naissances dans les très petites villes. En effet, on peut lire dans le Tableau 3 que 88% des individus nés à l'aube du XIX^e siècle sont nés dans une très petite commune. Leurs descendants sont également majoritairement nés dans une très petite commune, mais l'on note que plus les générations avancent, plus la part de naissance dans les moyennes et grandes villes augmente. Le déclin observé dans les données à partir des petits-enfants, soit dans le début de la seconde moitié du XIX^e siècle coïncide avec la période d'amorce de l'exode rural retenue par Lemercier et Rosental (2000). La faiblesse mesurée du déclin semble par ailleurs faire écho avec ce que nous rappellent Fine (1991, p. 93), à savoir que l'exode rural s'intensifie plutôt au XX^e siècle.

Reste à savoir l'origine des changements dans la proportion d'individus à naître dans chaque catégorie de ville. Pour apporter un élément de réponse, nous proposons de construire pour chaque génération, une matrice de transition conditionnelle au lieu de

Tableau 3. Répartition des lieux de naissance des individus de l'échantillon en fonction de la taille de la ville.

Type de commune	Âïeux	Enfants	Petits-enfants	Arrière-petits-enfants
Grande	3,40	3,92	6,31	7,63
Moyenne	5,18	5,25	6,55	7,86
Petite	3,43	3,56	3,73	3,81
Très petite	87,99	87,27	83,41	80,70

Note : Chaque colonne indique, pour une génération, le pourcentage d'individus de l'échantillon à naître dans chacune des trois catégories de villes.

Tableau 4. Déplacements transitoires conditionnels entre petites, moyennes et grandes villes, en pourcentages.

	Enfants				Petits-enfants				Arrière-petits-enfants			
	G	M	P	TP	G	M	P	TP	G	M	P	TP
G	67,10	4,14	2,42	26,34	52,67	7,47	2,93	36,92	36,73	9,97	3,27	50,03
M	5,75	67,97	1,96	24,32	10,42	47,99	3,23	38,36	14,02	33,85	3,02	49,11
P	4,63	4,11	65,37	25,89	10,21	7,64	37,46	44,69	12,50	11,13	21,38	54,99
TP	1,13	1,52	1,29	96,06	3,62	3,92	2,39	90,07	5,46	5,91	3,22	85,42

Note : Le tableau propose une matrice de transition conditionnelle pour chaque génération indiquant la fréquence de naissance des descendants dans une grande (G), moyenne (M), petite (P) ou très petite (TP) ville (en colonne), conditionnellement au lieu de naissance des aïeux (en ligne).

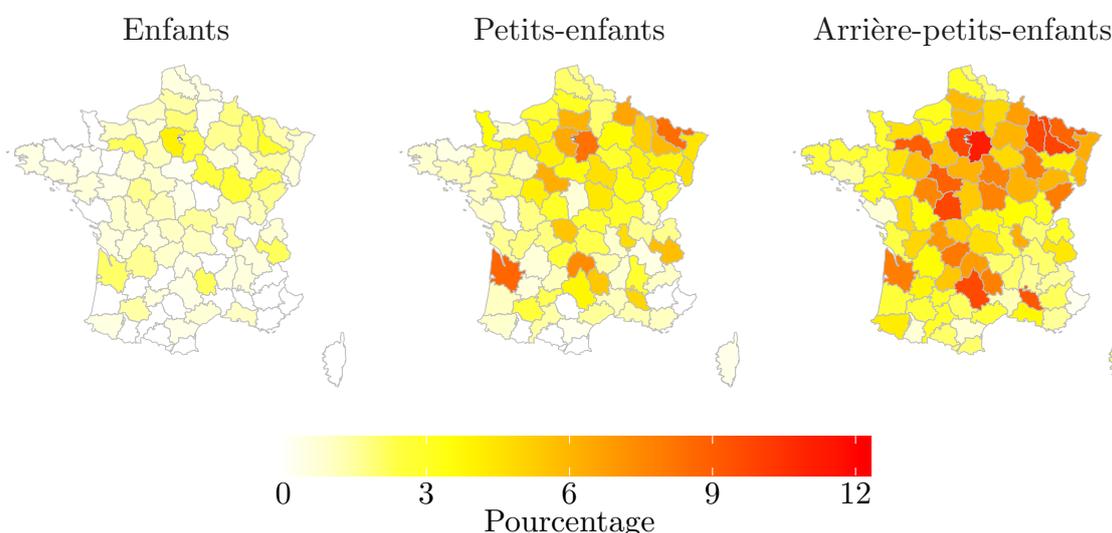
naissance des aïeux. Pour une génération donnée, nous classons les aïeux en fonction de la catégorie de ville dans laquelle ils sont nés. Puis, pour chaque catégorie de ville, nous repérons le lieu de naissance des descendants et calculons la proportion à naître dans chacune des quatre catégories de villes. Les résultats sont consignés dans le [Tableau 4](#). On note une écrasante majorité de sédentaires parmi les descendants des personnes nées dans les très petites villes. Toutefois, la proportion de sédentaires décline légèrement d'une génération à l'autre, passant de 96% pour les enfants à 85% pour les arrière-petits-enfants, au profit d'une compensation dans les moyennes et grandes villes principalement. Pour les aïeux nés dans des grandes et moyennes villes, le schéma est différent. On remarque certes une part importante de sédentaires parmi les descendants, mais relativement moins importante que celle des aïeux nés dans les très petites villes. Il ressort également un mouvement vers les très petites villes croissant au fur et à mesure des générations. En effet, la proportion de descendants, dont les aïeux sont nés dans une grande ville, à naître dans une très petite ville s'élève à 26% pour les enfants, et grimpe à 50% pour les arrière-petits-enfants. Il ne faut pas négliger que la majorité des individus naissent dans des très petites communes pour chaque génération. Cela ne vient pas contredire l'amorce de l'exode rural sur la seconde moitié du XIX^e siècle. Une migration de 5% des individus nés dans une petite ville vers une grande représente plus d'individus que 50% d'individus nés dans une grande ville.

5.3 Le cas parisien

Qu'en est-il de l'émigration parisienne ? Nous souhaitons regarder s'il existe des faits notables en relation avec l'émigration vers la capitale du pays.

Une première vision de la migration vers la capitale consiste à regarder, pour chaque département, le pourcentage des descendants nés à Paris. C'est ce que propose la [Figure 9](#)¹³. On note que pour les individus de la première génération, c'est-à-dire les enfants, le pourcentage de descendants nés à Paris est faible pour la plupart des départements. En moyenne, par département, un peu moins de 1% des enfants des individus nés entre 1800 et 1804 naissent à Paris. Il faut attendre la génération suivante pour commencer à apercevoir davantage de mouvements vers Paris. La moyenne départementale de descendants à Paris passe de 0,88% pour les enfants à 2,73% pour les petits-enfants. Certains départements se démarquent par des valeurs plus élevées, notamment la Gironde, pour laquelle 8,65% des petits enfants naissent à Paris. On relève également que les départements du sud et de l'ouest de la France se distinguent par des valeurs faibles. Enfin, pour la génération des arrière-petits-enfants, la moyenne départementale de naissances à Paris grimpe à 4,69%, avec un maximum de 11,2% pour la Seine-et-Marne.

FIGURE 9. Pourcentage de migrants vers Paris.



Note : Pour chaque génération (enfants à gauche, petits-enfants au milieu et arrière-petits-enfants à droite), et chaque département à l'exception de Paris, les cartes indiquent le pourcentage de descendants nés à Paris.

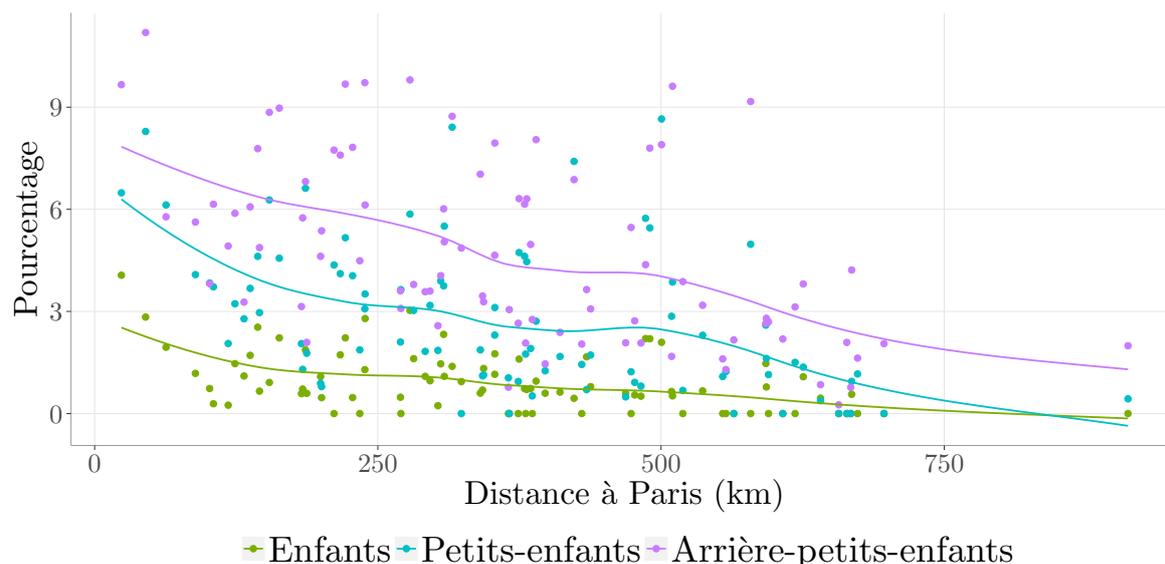
Il semble se dégager de ces cartes un effet distance : plus le département est proche de Paris, et plus la proportion de descendants à naître dans la capitale est élevée. C'est en effet ce que l'on peut observer à la [Figure 10](#), pour laquelle nous traçons le pourcentage de descendants à Paris en fonction de la distance des départements d'origine des aïeux à la capitale¹⁴. Les tendances négatives supportent bien l'idée que plus les départements sont proches de Paris, plus la proportion de descendants à y naître sont élevées, et ce, quelle que soit la génération. De plus, plus on avance dans les générations, plus les proportions moyennes sont importantes.

Pour poursuivre dans l'idée de l'effet de la distance sur les migrations à Paris, il est possible de regarder ce qu'il se passe à une échelle spatiale plus restreinte, en se penchant

13. On exclut le département de Paris pour des soucis visuels, pour que le fort pourcentage associé à ce département ne vienne pas gonfler l'échelle.

14. La distance entre les régions et Paris est calculée en utilisant la distance séparant les barycentres des régions.

FIGURE 10. Proportions départementales de descendants nés à Paris en fonction de la distance séparant chaque département à la capitale.



Note : Chaque point représente la proportion de descendants d'un département (en vert pour les enfants, en turquoise pour les petits-enfants et en mauve pour les arrière-petits-enfants) nés à Paris. La distance d'un département à la capitale retenue ici est celle qui sépare le barycentre de ce département à celui de la région de Paris. Les lignes correspondent à un lissage de type *loess*.

désormais sur les migrations sortantes de Paris. Nous regardons spécifiquement les déplacements transitoires entre Paris, sa couronne et le reste de la France. Nous définissons ici la couronne comme les alentours de Paris dans un rayon de 20km par rapport à ses limites. Nous reportons dans le [Tableau 5](#) des matrices de transition conditionnelles pour chaque génération, qui indiquent, conditionnellement au lieu de naissance des individus (à Paris ou dans sa Couronne), le lieu de naissance des descendants (à Paris, dans sa Couronne ou ailleurs). Lorsque les aïeux sont nés à Paris, la proportion de leurs descendants à naître également à Paris est importante, mais décroît de génération en génération, passant de 79,79% des enfants à 43,45% des arrière-petits-enfants. Tandis qu'une part croissante des descendants naît dans la couronne de Paris, une part également croissante, mais relativement plus importante naît au-delà de la couronne de 20km entourant Paris. Alors que l'on observe des mouvements progressifs de départs de la capitale, des flux opposés s'observent dans la couronne. Tandis que seuls 4,36% des enfants d'individus nés dans la couronne de Paris naissent dans la capitale, ce pourcentage grimpe les générations suivantes pour atteindre 11,79% pour les arrière-petits-enfants.

Tableau 5. Déplacements transitoires conditionnels entre Paris et sa couronne.

		Enfants			Petits-enfants			Arrière-petits-enfants		
		Paris	Couronne	Autre	Paris	Couronne	Autre	Paris	Couronne	Autre
Aïeul	Paris	74,79	3,56	21,64	55,04	9,07	35,89	43,45	8,98	47,58
	Couronne	4,36	79,65	15,99	6,77	68,42	24,81	11,79	58,19	30,02

Note : Le tableau propose une matrice de transition conditionnelle pour chaque génération indiquant la fréquence de naissance des descendants à Paris, dans sa couronne de 20km, ou plus loin (en colonne) conditionnellement au lieu de naissance des aïeux (en ligne).

Références

- BEAN, L. L., MAY, D. L. et 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](https://doi.org/10.1080/01615440.1978.9955216).
- BLAYO, Y. (1975a). La mortalité en France de 1740 a 1829. *Population*, 30:123. doi :[10.2307/1530647](https://doi.org/10.2307/1530647).
- BLAYO, Y. (1975b). Mouvement naturel de la population française de 1740 a 1829. *Population*, 30:15. doi :[10.2307/1530644](https://doi.org/10.2307/1530644).
- BLAYO, Y. et HENRY, L. (1967). Données démographiques sur la Bretagne et l'Anjou de 1740 à 1829. *Annales de démographie historique*, 1967(1):91–171. doi :[10.3406/adh.1967.955](https://doi.org/10.3406/adh.1967.955).
- BONNEUIL, N., BRINGÉ, A. et 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](https://doi.org/10.1080/03071020701833325).
- BOURDIEU, J., KESZTENBAUM, L. et POSTEL-VINAY, G. (2014). L'enquête TRA, une matrice d'histoire. *Population*, 69(2):217–248. doi :[10.3917/popu.1402.0217](https://doi.org/10.3917/popu.1402.0217).
- BOURDIEU, J., POSTEL-VINAY, G., ROSENTAL, P.-A. et SUWA-EISENMANN, A. (2000). Migrations et transmissions inter-générationnelles dans la France du XIXe et du début du XXe siècle. *Annales. Histoire, Sciences Sociales*, 55(4):749–789. doi :[10.3406/ahess.2000.279879](https://doi.org/10.3406/ahess.2000.279879).
- BOURDIEU, J., POSTEL-VINAY, G., ROSENTAL, P.-A. et SUWA-EISENMANN, A. (2004). La dispersion spatiale des familles : un problème de taille. les solidarités familiales de 1800 à 1940. *Recherches et prévisions*, 77(1):63–72. doi :[10.3406/caf.2004.2806](https://doi.org/10.3406/caf.2004.2806).
- BRUNET, G. et BIDEAU, A. (2001). Démographie historique et généalogie. *Annales de démographie historique*, 2000(2):101–110. doi :[10.3406/adh.2001.1977](https://doi.org/10.3406/adh.2001.1977).
- BRUNET, G. et VÉZINA, H. (2015). Les approches intergénérationnelles en démographie historique. *Annales de démographie historique*, 129(1):77. doi :[10.3917/adh.129.0077](https://doi.org/10.3917/adh.129.0077).
- CHESNAIS, J. (1986). *La Transition démographique : etapes, formes, implications économiques*. Travaux et documents - Institut national d'études démographiques. Presses universitaires de France.
- COHEN, W., RAVIKUMAR, P. et 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](https://doi.org/10.1017/s0022050717000468).
- DRIBE, M. et LUNDH, C. (2005). People on the move : determinants of servant migration in nineteenth-century Sweden. *Continuity and Change*, 20(1):53–91. doi :[10.1017/s026841600400534x](https://doi.org/10.1017/s026841600400534x).

- DUPÂQUIER, J. (1981). Une grande enquête sur la mobilité géographique et sociale aux XIXe et XXe siècles. *Population*, 36(6):1164–1167. doi :[10.2307/1532329](https://doi.org/10.2307/1532329).
- FINE, A. (1991). *La population française au XIXe siècle*, volume 1420. Presses Universitaires de France-PUF.
- FIRE, M. et 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](https://doi.org/10.1145/2700464).
- FLEURY, M. et HENRY, L. (1958). Pour connaître la population de la France depuis Louis XIV. plan de travaux par sondage. *Population*, 13(4):663. doi :[10.2307/1525088](https://doi.org/10.2307/1525088).
- GAVRILOV, L. A. et GAVRILOVA, N. S. (2001). Etude biodémographique des déterminants familiaux de la longévité humaine. *Population*, 56(1/2):225. doi :[10.2307/1534823](https://doi.org/10.2307/1534823).
- GAVRILOV, L. A., GAVRILOVA, N. S., OLSHANSKY, S. J. et CARNES, B. A. (2002). Genealogical data and the biodémography of human longevity. *Social Biology*, 49(3-4):160–173. doi :[10.1080/19485565.2002.9989056](https://doi.org/10.1080/19485565.2002.9989056).
- GAVRILOVA, N. S. et 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](https://doi.org/10.1080/10920277.2007.10597437).
- GIRRES, J.-F. et TOUYA, G. (2010). Quality assessment of the french OpenStreetMap dataset. *Transactions in GIS*, 14(4):435–459. doi :[10.1111/j.1467-9671.2010.01203.x](https://doi.org/10.1111/j.1467-9671.2010.01203.x).
- GRIFFITH, M., SPIES, N. C., KRYSIAK, K., MCMICHAEL, J. F., COFFMAN, A. C., DANOS, A. M., AINSCOUGH, B. J., RAMIREZ, C. A., RIEKE, D. T., KUJAN, L., BARNELL, E. K., WAGNER, A. H., SKIDMORE, Z. L., WOLLAM, A., LIU, C. J., JONES, M. R., BILSKI, R. L., LESURF, R., FENG, Y.-Y., SHAH, N. M., BONAKDAR, M., TRANI, L., MATLOCK, M., RAMU, A., CAMPBELL, K. M., SPIES, G. C., GRAUBERT, A. P., GANGAVARAPU, K., ELDRED, J. M., LARSON, D. E., WALKER, J. R., GOOD, B. M., WU, C., SU, A. I., DIENSTMANN, R., MARGOLIN, A. A., TAMBORERO, D., LOPEZ-BIGAS, N., JONES, S. J. M., BOSE, R., SPENCER, D. H., WARTMAN, L. D., WILSON, R. K., MARDIS, E. R. et GRIFFITH, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174. doi :[10.1038/ng.3774](https://doi.org/10.1038/ng.3774).
- HAKLAY, M. (2010). How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B : Planning and Design*, 37(4):682–703. doi :[10.1068/b35097](https://doi.org/10.1068/b35097).
- HENRY, L. (1956). Anciennes familles genevoises. étude démographique : XVIe - XXe siècle. *Population*, 11(2):334. doi :[10.2307/1524668](https://doi.org/10.2307/1524668).
- HENRY, L. et BLAYO, Y. (1975). La population de la France de 1740 à 1860. *Population (French Edition)*, 30:71. doi :[10.2307/1530646](https://doi.org/10.2307/1530646).
- HO., J. (1971). Les migrations intérieures en France à la fin du XVIIIe et au début du XIXe siècle. *Population (French Edition)*, 26(4):743. doi :[10.2307/1530646](https://doi.org/10.2307/1530646).

- HOUDAILLE, J. et TUGAULT, Y. (1987). Une bourgeoisie peu malthusienne dans un pays neuf : généalogies américaines du XIXe siècle. *Population (French Edition)*, 42(2):305. doi :[10.2307/1533086](https://doi.org/10.2307/1533086).
- 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. et ERLICH, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*. doi :[10.1126/science.aam9309](https://doi.org/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](https://doi.org/10.1016/j.hisfam.2008.01.006).
- KESZTENBAUM, L. (2014). L'étude des migrations grâce aux registres matricules militaires. *Popolazione e storia*, 14(2):9–38.
- LEASE, M. et YILMAZ, E. (2013). Crowdsourcing for information retrieval : introduction to the special issue. *Information Retrieval*, 16(2):91–100. doi :[10.1007/s10791-013-9222-7](https://doi.org/10.1007/s10791-013-9222-7).
- LEMERCIER, C. et ROSENTAL, P.-A. (2000). "pays" ruraux et découpage de l'espace : les réseaux migratoires dans la région lilloise au milieu du xixe siècle. *Population*, 55(4/5):691–725. doi :[10.2307/1534691](https://doi.org/10.2307/1534691).
- LINDAHL-JACOBSEN, R., HANSON, H. A., OKSUZYAN, A., MINEAU, G. P., CHRISTENSEN, K. et 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](https://doi.org/10.1016/j.annepidem.2013.02.001).
- MATTHIJS, K. et 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](https://doi.org/10.1016/j.hisfam.2010.01.002).
- MOTTA, M., MALAGUARNERA, M., FERRARI, E., MAURO, V. N., FERRUCCI, L., RAPISARDA, R., TOMASELLO, F., BASILE, G., FERLITO, L., PASSAMONTE, M. *et al.* (2007). Genealogy of centenarians and their relatives : a study of 12 families. *Archives of gerontology and geriatrics*, 45(1):97–102. doi :[10.1016/j.archger.2006.10.004](https://doi.org/10.1016/j.archger.2006.10.004).
- PERLS, T. T., WILMOTH, J., LEVENSON, R., DRINKWATER, M., COHEN, M., BOGAN, H., JOYCE, E., BREWSTER, S., KUNKEL, L. et PUCA, A. (2002). Life-long sustained mortality advantage of siblings of centenarians. *Proceedings of the National Academy of Sciences*, 99(12):8442–8447. doi :<https://doi.org/10.1073/pnas.122587599>.
- ROSENTAL, P.-A. (2004). La migration des femmes (et des hommes) en France au XIXe siècle. *Annales de démographie historique*, 107(1):107. doi :[10.3917/adh.107.0107](https://doi.org/10.3917/adh.107.0107).
- STATISTIQUE GÉNÉRALE DE LA FRANCE (2010). Données sur la démographie, la population et l'enseignement primaire sur la période 1800-1925. <https://www.insee.fr/fr/statistiques/2591293?sommaire=2591397>. Consulté le 9 février 2018.
- TOULEMON, L. (1995). Très peu de couples restent volontairement sans enfant. *Population*, 50(4/5):1079. doi :[10.2307/1534316](https://doi.org/10.2307/1534316).

- VALLIN, J., national d'études démographiques (FRANCE), I. et MESLÉ, F. (2001). *Tables de mortalité françaises pour les XIXe et XXe siècles et projections pour le XXIe siècle*. Éditions de l'Institut national d'études démographiques.
- van de WALLE, E. (1973). La mortalité des départements français ruraux au XIXe siècle. *Annales de démographie historique*, 1973(1):581–589. doi :[10.3406/adh.1973.1164](https://doi.org/10.3406/adh.1973.1164).
- van LEEUWEN, M. H., MAAS, I., RÉBAUDO, D. et PÉLISSIER, J.-P. (2015). Social mobility in France 1720–1986 : Effects of wars, revolution and economic change. *Journal of Social History*, 49(3):585–616. doi :[10.1093/jsh/shv054](https://doi.org/10.1093/jsh/shv054).
- VANDESCHRICK, C. (1992). Le diagramme de lexis revisité. *Population (French Edition)*, 47(5):1241. doi :[10.2307/1533940](https://doi.org/10.2307/1533940).
- WARBY, S. C., WENDT, S. L., WELINDER, P., MUNK, E. G. S., CARRILLO, O., SORENSEN, H. B. D., JENNUM, P., PEPPARD, P. E., PERONA, P. et MIGNOT, E. (2014). Sleep-spindle detection : crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods*, 11(4):385–392. doi :[10.1038/nmeth.2855](https://doi.org/10.1038/nmeth.2855).

A Données

Tableau 6. Déplacements intergénérationnels en France en fonction du genre.

Distance (km)	Enfants				Petits-enfants				Arrière-petits-enfants			
	Femmes		Hommes		Femmes		Hommes		Femmes		Hommes	
	%	(% cum.)	%	(% cum.)	%	(% cum.)	%	(% cum.)	%	(% cum.)	%	(% cum.)
[0,1]	62,24	(62,24)	63,74	(63,74)	38,54	(38,54)	39,07	(39,07)	24,32	(24,32)	25,21	(25,21)
(1,2]	1,55	(63,78)	1,82	(65,55)	1,88	(40,42)	1,88	(40,94)	1,64	(25,96)	1,82	(27,03)
(2,3]	2,26	(66,04)	2,15	(67,7)	2,66	(43,08)	2,71	(43,66)	2,2	(28,16)	2,26	(29,29)
(3,4]	2,15	(68,19)	2,41	(70,11)	3,19	(46,27)	2,89	(46,55)	2,72	(30,88)	2,66	(31,95)
(4,5]	2,1	(70,28)	2,08	(72,19)	2,7	(48,97)	2,67	(49,21)	2,59	(33,47)	2,39	(34,33)
(5,6]	1,88	(72,16)	1,51	(73,7)	2,39	(51,35)	2,57	(51,78)	2,38	(35,85)	2,34	(36,67)
(6,7]	1,5	(73,66)	1,37	(75,07)	2,18	(53,54)	1,91	(53,69)	2,04	(37,89)	2,05	(38,72)
(7,8]	1,24	(74,9)	1,25	(76,31)	1,75	(55,28)	1,61	(55,29)	1,62	(39,52)	1,8	(40,52)
(8,9]	1,09	(75,99)	0,92	(77,23)	1,54	(56,82)	1,46	(56,76)	1,55	(41,07)	1,6	(42,12)
(9,10]	1,11	(77,1)	0,82	(78,05)	1,53	(58,36)	1,3	(58,05)	1,29	(42,36)	1,47	(43,59)
(10,11]	0,78	(77,88)	0,64	(78,69)	1,14	(59,5)	1,16	(59,21)	1,34	(43,7)	1,29	(44,88)
(11,12]	0,49	(78,36)	0,69	(79,39)	0,97	(60,47)	0,91	(60,12)	1,01	(44,71)	1,11	(45,99)
(12,13]	0,52	(78,88)	0,49	(79,87)	0,93	(61,4)	0,98	(61,1)	1,03	(45,74)	1,14	(47,13)
(13,14]	0,42	(79,3)	0,35	(80,22)	0,95	(62,35)	0,8	(61,9)	0,89	(46,63)	0,84	(47,97)
(14,15]	0,44	(79,74)	0,27	(80,49)	0,79	(63,14)	0,67	(62,57)	0,81	(47,44)	0,85	(48,82)
(15,16]	0,39	(80,13)	0,43	(80,92)	0,6	(63,74)	0,8	(63,37)	0,78	(48,22)	0,89	(49,71)
(16,17]	0,38	(80,51)	0,33	(81,25)	0,73	(64,47)	0,77	(64,14)	0,68	(48,9)	0,69	(50,4)
(17,18]	0,32	(80,83)	0,35	(81,6)	0,63	(65,11)	0,57	(64,71)	0,73	(49,63)	0,66	(51,07)
(18,19]	0,31	(81,14)	0,26	(81,86)	0,67	(65,78)	0,5	(65,21)	0,69	(50,31)	0,65	(51,71)
(19,20]	0,32	(81,45)	0,27	(82,13)	0,5	(66,28)	0,49	(65,69)	0,57	(50,88)	0,64	(52,35)
(20,21]	0,22	(81,67)	0,26	(82,39)	0,43	(66,71)	0,45	(66,14)	0,55	(51,43)	0,5	(52,85)
(21,22]	0,25	(81,92)	0,24	(82,64)	0,35	(67,06)	0,46	(66,6)	0,52	(51,95)	0,44	(53,29)
(22,23]	0,28	(82,19)	0,24	(82,88)	0,34	(67,4)	0,34	(66,94)	0,54	(52,49)	0,41	(53,7)
(23,24]	0,29	(82,48)	0,16	(83,04)	0,33	(67,73)	0,46	(67,4)	0,4	(52,89)	0,43	(54,13)
(24,25]	0,19	(82,67)	0,09	(83,13)	0,25	(67,98)	0,28	(67,67)	0,32	(53,22)	0,35	(54,48)
(25,50]	2,51	(85,19)	3	(86,13)	5,07	(73,05)	5,36	(73,03)	7,07	(60,29)	6,97	(61,46)
(50,75]	1,57	(86,75)	1,27	(87,41)	2,84	(75,89)	2,7	(75,74)	3,68	(63,97)	3,95	(65,41)
(75,100]	1	(87,75)	1,11	(88,51)	1,84	(77,72)	1,82	(77,56)	2,63	(66,61)	2,71	(68,11)
(100,150]	1,72	(89,47)	1,59	(90,1)	2,83	(80,55)	2,78	(80,34)	4,23	(70,83)	4,08	(72,19)
(150,200]	1,48	(90,95)	1,34	(91,44)	2,82	(83,38)	2,49	(82,83)	3,75	(74,58)	3,5	(75,7)
(200,Inf]	9,05	(100)	8,56	(100)	16,62	(100)	17,17	(100)	25,42	(100)	24,3	(100)

Davantage d'informations relatives aux données sont présentées sur l'annexe en ligne : https://3wen.github.io/genealogie_fr/.