amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

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

Arthur Charpentier
Ewen Gallic

WP 2019 - Nr 16

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

Arthur Charpentier[a,b] & Ewen Gallic[1,c]

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

Mai 2019

A l'ère du numérique, les données peuvent être collectées massivement, de manière collaborative et à moindre coût. 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, utilisons les enregistrements concernant 2 457 450 individus français ou d'origine française ayant vécu au XIX^{e} siècle. Dans un premier temps, nous étudions la qualité de ces données. Nous mettons en évidence la présence de biais importants, notamment concernant le genre des individus. Les femmes sont sous-représentées dans les données comparativement aux hommes. Des biais relatifs à la fécondité sont également observés. En dépit de ces limites dont souffrent les données collaboratives de généalogie, nous montrons dans un deuxième temps qu'il est possible de retrouver des résultats connus dans la littérature en démographie historique. Plus particulièrement, nous exploitons les dates de naissance et de décès afin d'examiner la mortalité des individus présents dans la base de données. Nous exploitons également la richesse des caractéristiques spatiales contenues dans les arbres généalogiques pour analyser les migrations internes en France.

Mots-clés : Généalogie, Données collaboratives, Longévité, Migration, XIXe siècle, R.

1. Introduction et motivations

Au cœur de la démographie historique se trouvent les données. Leur provenance 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 utilisés (Henry & Blayo, 1975; Dupâquier, 1981; van de Walle, 1986; Bourdieu, Postel-Vinay, Rosental, & Suwa-Eisenmann, 2004; Bonneuil, Bringé, & Rosental, 2008), mais ne constituent pas l'unique source présente dans la littérature. Les registres matricules militaires établis par l'administration militaire sont une autre source d'information, utilisée notamment pour étudier les migrations au cours de la vie (Ho, 1971; Kesztenbaum, 2008, 2014).

Les données de registres ne sont pas uniquement exploitées par les historiens et les démographes. Les généalogistes s'en saisissent également pour l'amélioration des connaissances de notre histoire. À titre d'exemple, les mormons ont constitué à l'Université

[1] Auteur correspondant. *Correspondance* : École d'Économie d'Aix-Marseille, Aix-Marseille Université, 5-9 Boulevard Bourdet, CS 50498, 13205 Marseille Cedex 1, France. *Courriel* : ewen.gallic@univ-amu.fr

d'Utah de volumineuses bases de données généalogiques ayant aidé à l'étude de la population dans le passé (Bean, May, & Skolnick, 1978; Lindahl-Jacobsen *et al.*, 2013). Toutefois, comme le note Dupâquier (1993), bien que le travail accomplis par les généalogistes rende « *de signalés services aux historiens* », il souffrirait de deux biais principaux. Le premier est lié à la représentativité des généalogistes, constituant un petit monde à part où la majorité des individus s'adonnant à cette activité est constituée d'hommes assez âgés, retraités. Le deuxième biais provient de la manière généralement ascendante par laquelle les généalogistes constituent leur reconstitution des familles. En procédant ainsi, c'est-à-dire en partant d'un individu et en remontant progressivement ses aïeux, les collatéraux n'ayant pas eu d'enfants tendent à être peu reportés. En conséquence, les milieux peu féconds ou inféconds tendent à ne pas être correctement représentés (Dupâquier et Blanchet, 1992).

L'essor d'Internet dans les années 2000 pourrait avoir modifié de manière considérable ces deux sources de biais. Tout d'abord, grâce à l'accès facilité aux fonds d'archives numérisés, permettant, comme l'indique Hervis (2012), à une population plus jeune et active de se livrer à des activités de généalogie. La démocratisation de ces activités permettrait donc de réduire le biais de représentativité. Hervis (2012) rappelle néanmoins que la plupart des généalogistes a plus de 50 ans. Ensuite, le développement d'Internet depuis les années 2000 a été accompagné de l'utilisation de plus en plus accrue de sites web de généalogie. Certains d'entre eux proposent des accès libres et gratuits aux registres numérisés, et invitent les internautes à les indexer de manière collaborative. Une fois indexés, les registres permettent aux utilisateurs de reconstituer aisément leur arbre généalogique et de le partager avec les autres utilisateurs. Ce partage permet potentiellement de reconstituer des branches collatérales plus aisément, venant diminuer les biais de représentativité des milieux inféconds.

Bien que la démarche adoptée par les sites de généalogie et leurs utilisateurs ne soit pas aussi rigoureuse que celle employée dans l'enquête des 3 000 familles initiée par Jacques Dupâquier au début des années 1980 (voir Dupâquier & Kessler, 1992 pour plus de détails sur l'enquête des 3 000 familles), elle offre quelques avantages non négligeables. La collaboration des internautes permet non seulement de couvrir un large nombre de familles, sur une échelle spatiale large, mais également de réunir les informations dans des temps relativement courts, et ce pour des coûts moindres.

De manière générale, les données collaboratives semblent offrir de prometteuses perspectives. D'un point de vue technique, elles fournissent un échantillon d'apprentissage permettant d'estimer les paramètres de modèles statistiques (Lease & 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 CIViC[2] visant à accroitre 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.*, 2014). En géographie, les données collaboratives, notamment celles d'Open Street Map[3] (OSM), le projet international visant à créer une carte libre du monde, sont aussi utilisées par des chercheurs. Haklay (2010) a 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 & Touya (2010) sont arrivés aux mêmes conclusions, en soulignant toutefois des difficultés liées à l'hétérogénéité des données.

Qu'en est-il des données collaboratives de généalogie ? Les questions de représentativité pointées par Dupâquier (1993) persistent-t-elles ? La démographie historique peut-elle tirer profit des données de généalogie renseignées par des millions d'utilisateurs ? Pour l'heure, ces questions ont été éludées lors de l'utilisation de ces données à des fins de recherche. Quelques travaux ont été menés à partir des données mises en commun par les utilisateurs des sites de généalogie tels wikitree.com, familysearch.org ou encore de geni.com. Ces travaux ont permis

[2] Clinical Interpretations of Variants in Cancer : https://civicdb.org/home.
[3] http://openstreetmap.fr/.

d'étudier la longévité des individus (Gavrilova & Gavrilov, 2007; Gavrilov, Gavrilova, Olshansky, & Carnes, 2002; Fire & Elovici, 2015). La récente étude de Kaplanis *et al.* (2018), qui explore les arbres généalogiques de plusieurs millions d'individus montre que les données de généalogie issues de la collaboration d'amateurs permettent d'obtenir des arbres généalogiques de grande qualité. Pour aller plus loin, à l'aide de données de généalogie issues du site du site web Geneanet[4], nous nous concentrons sur le cas français, en fournissant de plus amples comparaisons avec les résultats déjà connus de la littérature. Nous commençons par une partie méthodologique visant à expliquer les méthodes employées pour extraire et classer les informations contenues dans les données de Geneanet (Section **Erreur ! Source du renvoi introuvable.**). Nous explorons ensuite la représentativité des données (Section 3), puis examinons la mortalité des individus nés en France métropolitaine au début du XIXe siècle (Section 4) et leur migration ainsi que celle de leurs descendants sur trois générations (Section 5).

2. Construction des données

2.1. Périmètre des données

Ce travail s'appuie sur des données collectées par des amateurs de généalogie à la recherche de leurs ancêtres. 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. 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 les données sur un ensemble géographique plus large. Ici, les données dont nous disposons permettent de considérer le territoire de la France métropolitaine telle que définie aujourd'hui[5]. Par ailleurs, nous ne sommes pas forcés de 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 en France (voir Bourdieu *et al.* (2014) pour plus de détails) ou de la base COR en Belgique (Matthijs & Moreels, 2010). Nous pouvons suivre un nombre plus important d'individus, et ce sur 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 certaines 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

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

[5] 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. Plus précisément, notre étude s'appuie sur des individus nés en France métropolitaine entre 1800 et 1805 et leur descendants, ces derniers pouvant être nés hors du territoire français.

résulte de la taille des bases de 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. Il s'agit en effet 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[6]. Le second frein au regroupement est plus délicat. Il concerne la source 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 et présentent de nombreux problèmes ne garantissant pas *a priori* leur pertinence dès lors qu'il s'agit de les utiliser à des fins de recherche. La présence d'un individu et la quantité d'information à son sujet dans la base dépendant de l'existence d'une source, et à nouveau du 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. Une liste plus exhaustive des sources de problèmes liés aux données de généalogie est dressée dans l'article de Dupâquier (1993). Toutefois, la quantité d'informations disponible grâce au développement du numérique suggère que les efforts consentis par les généalogistes amateurs peuvent bénéficier aux historiens et démographes.

2.2. Construction des arbres généalogiques

Dans le cadre de cette étude, nous nous intéressons aux 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 descendants jusqu'à trois générations[7], soit 2 457 450 personnes. Cette période initiale correspond, comme le rappellent Fleury & Henry (1958), à la reprise sans interruption des relevés de mariages, naissance et décès dans l'ensemble de la France.

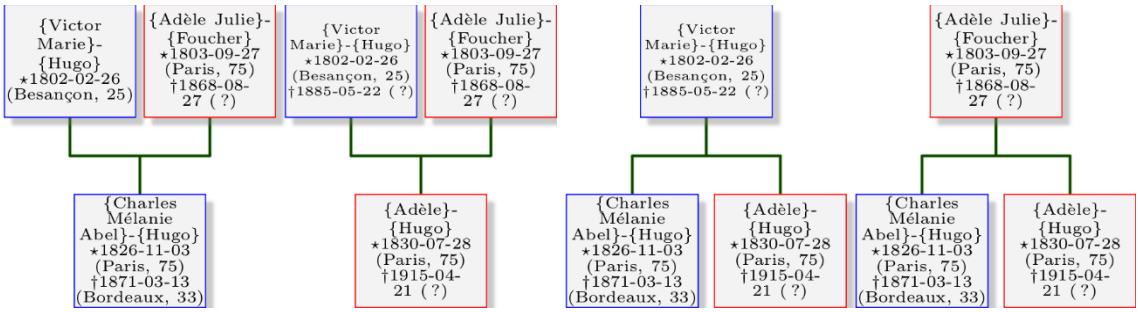
Typiquement, les généalogistes amateurs pratiquent principalement la généalogie ascendante pour construire une généalogie individuelle (Brunet & Vézina, 2015). C'est-à-dire qu'ils partent d'un individu et recherchent progressivement les aïeux. Nos données correspondent à ce type de généalogie : nous disposons d'informations directes concernant les parents – lorsque celles-ci sont disponibles – pour chaque individu présent dans la base. Or, lorsqu'il s'agit d'étudier la dispersion géographique à partir d'une poignée de femmes et d'hommes, comme le propose la section 5, 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 partie gauche de la Figure 1 illustre ce cas. Ce que nous souhaitons *in*

[6] Davantage d'informations relatives aux données sont présentées dans le document en ligne : https://3wen.github.io/genealogie_fr/.

[7] 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 individus de la génération 1800-1804, 402 190 enfants de ces individus, 286 071 petits-enfants et 222 103 arrière-petits-enfants, soit un total de 2 457 450 personnes).

fine, est de pouvoir suivre la descendance de Victor Hugo et d'Adèle Foucher, comme illustré sur la partie droite de la Figure 1[8].

Figure 1. Exemple d'un extrait d'arbre généalogique d'un utilisateur : à gauche : données telles qu'elles se présentent dans la base (deux individus avec les mêmes parents) ; à droite : données telles que l'on souhaite structurer (les descendants de chaque parent).



Note : Chaque rectangle représente une personne, les liens entre les rectangles représentent la filiation.

Un important travail de couplage s'avère donc 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, nous devons procéder à un travail important de nettoyage et de mise en forme des données[9]. 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[10]. La présence d'un même individu dans plusieurs arbres d'utilisateurs s'avère utile 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.

Pour repérer les doublons, nous appliquons de manière successive plusieurs méthodes. La première est la plus simple. Elle consiste à fusionner 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 ({Ma**tt**hieu Paul}-{Henri} ou {Ma**t**hieu Paul}-{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 prénoms[11]. 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 sur les dates incomplètes (jours ou

[8] On note ici qu'il s'agit d'un extrait de l'arbre d'un utilisateur. Par souci de simplicité, les frères et sœurs de Charles et Adèle Hugo ne sont pas présents sur la Figure 1.

[9] De plus amples informations sur la procédure peuvent être obtenues dans le document en ligne : https://3wen.github.io/genealogie_fr/.

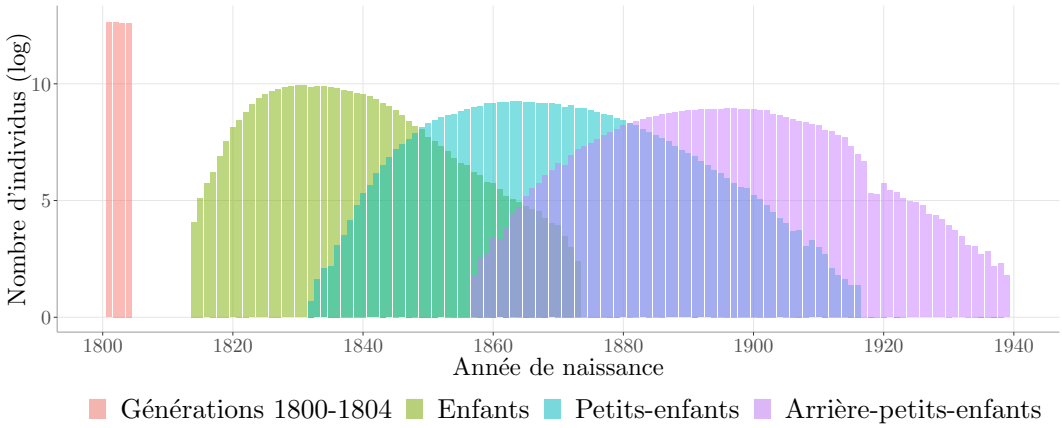
[10] Le découpage par département s'effectue en se basant sur le département de naissance des individus des générations 1800-1804. À l'issue des traitements par département, tous les arbres sont regroupés dans une même base et les doublons transdépartementaux sont recherchés et supprimés. Le document en ligne à l'adresse https://3wen.github.io/genealogie_fr/ fournit davantage d'explications et propose quelques statistiques et exemples.

[11] 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, Ravikumar, & Fienberg (2003).

mois manquants ou erronés). La cinquième façon d'identifier des doublons 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 doublons 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 individus des générations 1800-1804 ; (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.

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



Note : Ce graphique montre la distribution des années de naissances pour les individus nés en France métropolitaine entre 1800 et 1804, et pour leurs descendants sur trois générations. Les individus des générations 1800-1804 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. Le nombre de naissances par année est indiqué sur l'axe des ordonnées, suivant une échelle logarithmique., le nombre d'aïeux étant relativement important comparativement à leurs descendants.

3. Représentativité des données

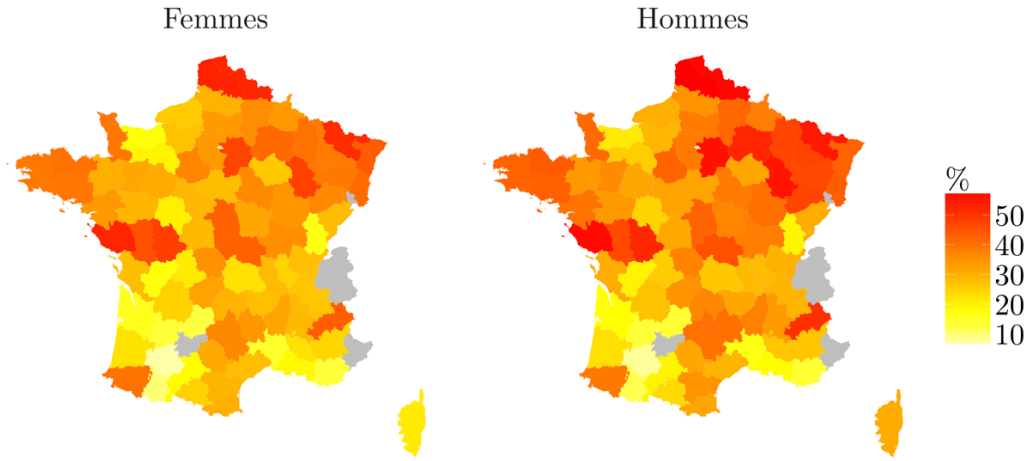
Pour évaluer l'intérêt que peuvent représenter les données collaboratives de généalogie pour les études démographiques, nous décrivons notre échantillon et le comparons avec des résultats provenant de la littérature.

3.1. Nombre de naissances par département

Une mesure de la représentativité géographique 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 nouveau-nés. Nous calculons le rapport entre les effectifs dans notre échantillon et les effectifs de l'Insee, selon le sexe (cartes de la Figure 3). La couverture varie selon les départements mais elle semble indépendante du sexe (la corrélation entre la couverture départementale des hommes avec celle des femmes est de 98%). Dans les départements situés dans le Nord et l'Est de la France, le nombre de naissances retrouvées dans

les données de Geneanet avoisine les 60% (plus précisément, pour les hommes, 57,7% dans le Pas-de-Calais, 57% dans le Nord, 54.5% en Moselle). Relativement, les départements situés dans le sud et le sud-ouest de la France affichent une proportion de naissances retrouvées dans les données bien moindre (à titre d'exemple, seulement 8,4% pour les hommes dans le Gers, 11,6% dans les Hautes-Pyrénées, 12,3% en Lot-et-Garonne). À l'échelle de la France métropolitaine, le pourcentage global des naissances retrouvées à l'aide des données de généalogie s'élève à 32,8%. La forte hétérogénéité observée entre les départements pourrait en partie être expliquée par l'accès plus ou moins tardif aux fonds d'archives publiques numérisées. Parmi les départements dans lesquels le pourcentage de naissances retrouvées dans les données de généalogie est le plus faible, figurent le Gard, le Gers et les Hautes-Pyrénées. Ces trois départements ont commencé à numériser leur état-civil ancien seulement récemment. La gratuité de l'accès aux données numérisées peut fournir un autre élément d'explication. On note par exemple un faible pourcentage de naissances retrouvées dans les données de Geneanet en Charente et dans le Calvados, deux départements dans lesquels l'accès gratuit aux archives numérisées n'est que très récent (2015 pour la Charente, 2016 pour le Calvados).

Figure 3. Proportion des naissances par département enregistrées par l'Insee retrouvées dans l'échantillon issu de Geneanet.



Note : Ces cartes montrent la couverture de l'échantillon par département en comparant le nombre d'individus nés en 1801 présents dans la base aux enregistrements de l'Insee. La couleur grise indique une valeur manquante. Plus la couleur vire au rouge, plus le nombre de naissances retrouvées dans les données de Geneanet est proche du chiffre officiel de l'Insee.

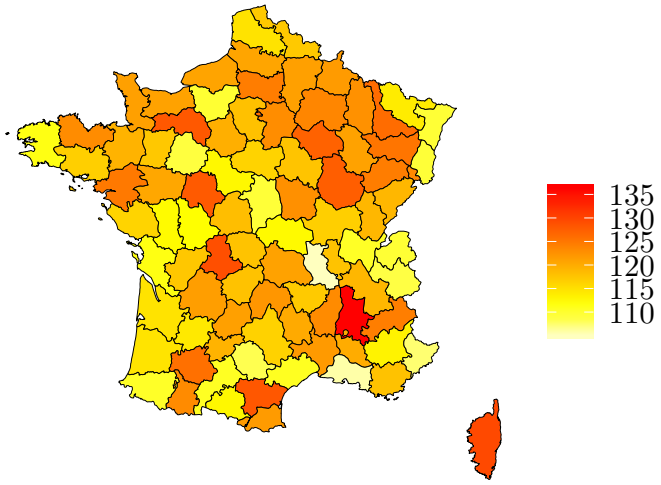
3.2. Rapport de masculinité

L'échantillon final comporte une forte surreprésentation des hommes. Le rapport de masculinité à la naissance, c'est-à-dire le rapport entre le nombre de naissances de garçons sur le nombre de naissances de filles, s'établit en effet à 116. En se concentrant uniquement sur les individus nés en 1801, nous obtenons un rapport de masculinité de 117[12]. Cette valeur est élevée au regard de ce que l'on trouve chez Blayo & Henry (1967), qui font état d'une valeur de 105,4 pour la période 1740–1829 pour la Bretagne et l'Anjou. Par ailleurs, les données de l'Insee

[12] En s'appuyant sur les taux d'échantillonnage par département, c'est-à-dire le rapport entre la taille de l'échantillon sur la taille de la population indiquée par l'Insee, nous pouvons calculer des intervalles de confiance pour encadrer cette valeur, en suivant la méthode décrite par Brian & Jaisson (2007). Nous obtenons des bornes de 116 et 120, pour un degré de confiance à 95%. Le Tableau 3 en annexe reporte l'ensemble des valeurs.

concernant les naissances par sexe indiquent, pour 1801, un rapport de masculinité de 105. Cette surreprésentation masculine s'observe pour l'ensemble des départements, avec un minimum de 105 (soit 51,2% d'hommes) dans la Loire et un maximum de 136 (soit 57,7% d'hommes) dans la Drôme[13] (Figure 4). On note une corrélation inverse forte entre les rapports de masculinité élevés et les taux d'échantillonnage.

Figure 4. Rapport de masculinité à la naissance par département, pour les individus nés en 1801.



Note : Les rapports de masculinité par département sont calculés à partir des observations relatives aux individus nés en 1801 uniquement.

L'existence d'un biais sexiste dans les études de généalogie a déjà été observée. Gavrilov & 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 relatives aux 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 lignées masculines, les épouses étant toutefois ajoutées à chaque génération. Cette technique de constitution de l'échantillon suivant les lignes patriarcales est d'ailleurs similaire à celle aboutissant aux données dont nous disposons. Brunet et Bideau (2001) notent que les généalogistes amateurs « *se limitent généralement à une lignée ascendante, avec une exploitation plus ou moins fouillée des branches collatérales* ». L'observation d'un biais sexiste dans notre étude ne vient pas contredire cette affirmation. D'un côté, il est vrai que techniquement, la maternité étant plus simple à établir que la paternité, il serait plus légitime de suivre des lignes matrilinéaires, comme le note Balsamo (1999). Cela dit, comme le rappelle Eichner (2014), en vertu du Code Napoléonien, « *les femmes sont subsumées par l'état civil de leur mari* » - par exemple, une femme prend la nationalité de son mari au moment du mariage. Ainsi, lors d'un mariage entre deux individus, suite au changement de nom de la femme, il devient peut-être plus difficile de suivre cette dernière sur les registres. Dans le cas où le registre de mariage n'existe plus ou n'est pas retrouvé par le généalogiste, la trace de la femme est potentiellement perdue.

[13] Les intervalles de confiance correspondant sont de 91 à 121 pour la Loire et de 116 à 160 pour la Drôme.

3.3. Fécondité

Dans l'échantillon, les femmes ayant été mariées au cours de leur vie et ayant atteint un âge d'au moins 15 ans[14] ont eu en moyenne 1,46 enfant. Pour les femmes de la génération initiale (1800-1804), 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). Ces niveaux sont bien inférieurs 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 & Bideau (2001), certains individus, notamment ceux restés célibataires, ne sont pas nécessairement intégrés. Hollingsworth (1976) explique par ailleurs que les généalogistes ne seraient pas vraiment intéressés par la question de la mortalité infantile dans le passé, l'effort à fournir pour retrouver si un enfant mentionné une fois est mort jeune ou pas étant trop conséquent. L'absence de ces individus impliquerait *de facto* une mesure biaisée du taux de fécondité. La distribution du nombre d'enfants par femme mariée en fonction de chaque génération est donnée dans le

Tableau 1. Si la part de femmes sans enfant pour les individus de la génération initiale (47%) apparaît très élevée au regard des proportions avancées par Houdaille & Tugault (1987) qui avoisinent plutôt les 15%, celles des générations suivantes (27% et 26%) 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).

	Nombre d'enfants												
Génération	0	1	2	3	4	5	6	7	8	9	10	>10	Total
1800-1804	46,9	24,5	10,3	6,1	4,0	2,6	1,9	1,2	0,8	0,5	0,4	0,6	100
Enfants	27,4	39,2	13,2	7,5	4,8	2,9	1,9	1,2	0,8	0,5	0,3	0,4	100
Petits-enfants	25,8	40,3	13,7	7,9	4,5	2,7	1,9	1,2	0,8	0,5	0,3	0,5	100

Note : ce tableau indique pour chaque génération (en lignes), la répartition du nombre d'enfants par femme mariée (en colonnes) dans les données issues de Geneanet. L'échantillon s'arrêtant aux arrière-petits-enfants des individus nés entre 1800 et 1804, il ne renseigne pas sur le nombre d'enfants que ces arrière-petits-enfants ont eux-mêmes eus.

4. Mortalité des français au début du XIXe siècle

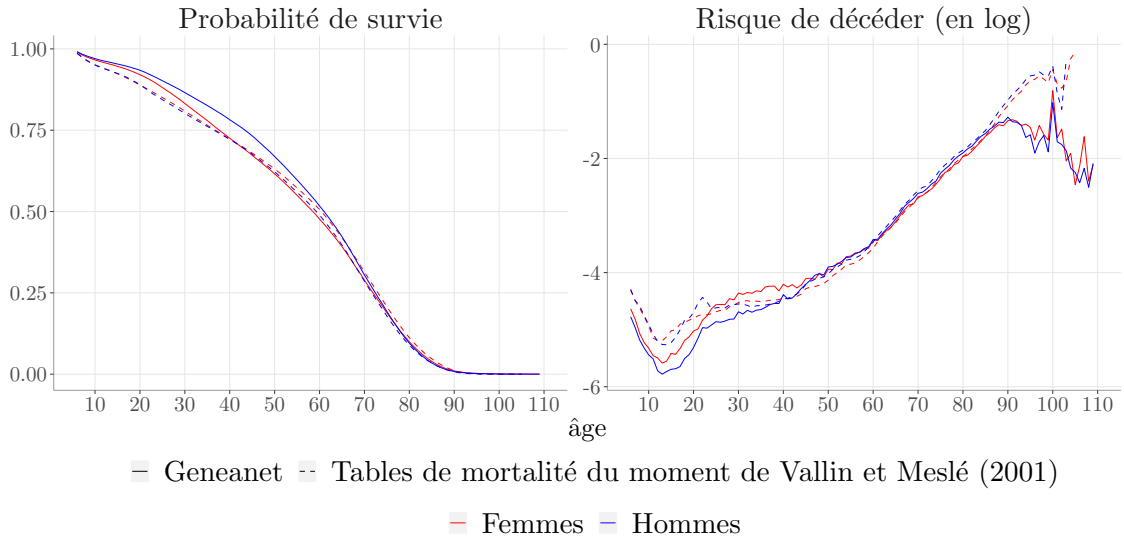
En combinant la date de naissance et celle de décès, une étude de la mortalité peut être réalisée. Cette section propose d'explorer les questions de mortalité et de survie d'individus, à l'échelle nationale puis régionale.

Les cohortes que nous décidons de suivre sont au nombre de 5. Il s'agit de celles des générations 1800 à 1804, incluses. Pour chaque cohorte, en utilisant les dates de naissance et de décès, qui sont indiquées pour 813 551 des individus de ces cohortes (soit 53% d'entre eux), nous calculons le nombre d'individus encore en vie ainsi que le nombre de décès à 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 par génération

[14] Soit l'âge minimum au mariage pour les femmes tel que précisé dans un décret du 17 mars 1803 (Henry & Houdaille, 1979).

proposées par Vallin & Meslé (2001)[15]. Ces tables commencent à partir de 1806, ce qui nous empêche *de facto* de comparer les probabilités de survie sur les premières années de vie des individus des générations 1800-1804 de notre base[16]. Aussi, pour chacune des cinq cohortes, nous estimons des probabilités de survie conditionnelles : nous considérons uniquement les individus ayant au moins atteint l'âge de 6 ans. Puis, nous agrégeons nos estimations : pour chaque âge. La probabilité de survie que nous reportons correspond à la moyenne des cinq valeurs estimées séparément pour chaque cohorte. Afin de rendre la comparaison cohérente avec les tables de Vallin et Meslé, nous appliquons la même méthode d'estimation : nous estimons la probabilité conditionnelle de survie à partir des tables de mortalité, pour chaque cohorte de 1800 à 1804 et agrégeons le résultat à l'aide d'une moyenne arithmétique.

Figure 5. Comparaison des courbes de survie et des risques de décéder selon l'âge pour les femmes et les hommes estimées à partir de Geneanet avec les estimations issues des tables de mortalité par générations de Vallin et Meslé (2001).



Note : Les estimations sont réalisées conditionnellement au fait que les individus aient survécu au moins jusqu'à leur sixième année d'existence.

Nos estimations réalisées à partir des données de Geneanet fournissent une surestimation de la survie des individus, particulièrement en ce qui concerne les hommes (Figure 5). Nous sous-estimons principalement la probabilité de décès des individus jeunes ainsi que celle des individus très âgés. Pour 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 individus des générations 1800-1804) l'estimation est bien meilleure.

En effectuant cette fois l'analyse sur l'ensemble des individus des cohortes 1800 à 1804, sans nous limiter à ceux ayant survécu au moins jusqu'à leur sixième année, nous pouvons calculer l'espérance de vie à la naissance de ces cohortes. Nous obtenons des valeurs de 41,8 années pour les femmes et 43,5 années pour les hommes. Les intervalles de confiance à 95% obtenus par *bootstrap*[17] encadrent ces espérances de vie à la naissance entre 41,6 et 42,0 années pour les femmes et 43,3 et 43,7 années pour les hommes.

Ces espérances de vie à la naissance sont supérieures à celles que l'on peut calculer avec les tables de mortalité de Vallin & Meslé (2001) pour la cohorte 1806, c'est-à-dire la première

[15] Les tables de mortalité de Vallin & Meslé (2001), sont disponibles en ligne à l'adresse suivante : https://www.ined.fr/Xtradocs/cdrom_vallin_mesle/texte.pdf (tableau III-A-1**).**

[16] Par la suite, nous calculons des espérances de vie à la naissance à partir des tables de mortalité complètes.

[17] Pour de plus amples détails sur la méthode, on consultera le document en ligne https://3wen.github.io/genealogie_fr/.

cohorte complète dans ces tables. En effet, les valeurs obtenues à partir des tables de mortalité sont de 37,16 pour les femmes avec un intervalle de confiance à 95% allant de 37,0 à 37,4 ; elles sont de 32,6 pour les hommes avec un intervalle de confiance à 95% allant de 32,5 à 32,8.

Les données de généalogie collaboratives souffrent donc d'un biais dans l'estimation de l'espérance de vie à la naissance. Ce biais pourrait s'expliquer par une sous-représentation des enfants décédés en bas âge dans les données généalogiques communautaires, comme mentionné par Gavrilov et Gavrilova (2001). Ces enfants décédés en bas-âge n'ayant pas connu de descendance, la trace de leur existence serait moins grande que celle d'individus ayant eu de nombreux enfants. Aussi, la probabilité de retrouver ces individus dans un arbre de généalogie en serait diminuée. Nous souhaitons savoir si, en revanche, les différences géographiques sont plus fidèlement retranscrites[18]. 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 à la naissance par département français. Un de ces regroupements concerne les cohortes de 1801 à 1810, sur lequel nous nous appuyons. Nous calculons, 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 pour chaque département, par rapport à la moyenne nationale. Les résultats sont proposés sur les cartes de la Figure 6. Globalement, les contrastes départementaux sont plutôt similaires entre les deux sources, la corrélation spatiale s'élevant à 0,65. 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.

5. Sédentarité et migration intérieure en France au XIXe siècle

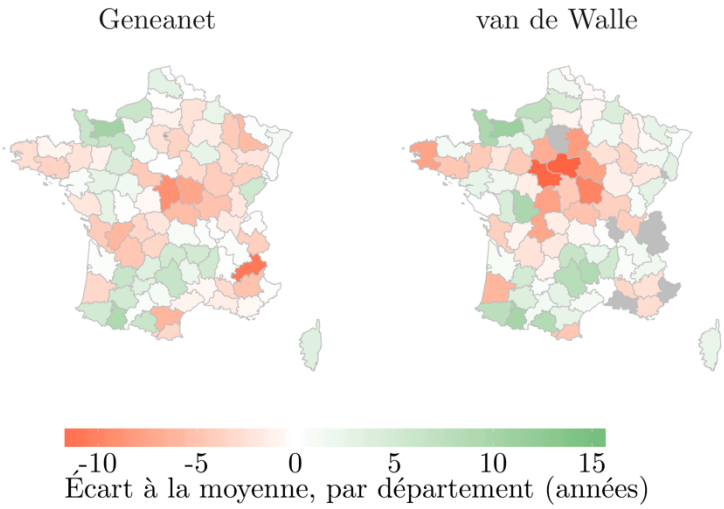
Outre les dates associées aux événements de naissance, mariage ou de décès, les données de généalogie fournissent une indication des lieux, sous forme de coordonnées GPS[19]. En suivant de génération en génération où naissent les individus, nous pouvons utiliser les données de généalogie afin d'étudier les migrations définitives de glissement, autrement dit les déplacements pérennes d'individus : les montagnards vont remplacer les gens des plaines partis en villes[20].

Figure 6. Écarts à la moyenne nationale de l'espérance de vie à la naissance par département pour les cohortes 1800 -1804 (Geneanet) et les cohortes de 1801-1810 (van de Walle).

[18] Des représentations graphiques de l'hétérogénéité spatiale des estimations sont proposées dans le document en ligne à l'adresse : https://3wen.github.io/genealogie_fr/.

[19] Lorsqu'il a été possible de les identifier, les lieux sont géolocalisés par Geneanet. Des couples de coordonnées (longitude, latitude) sont alors fournis pour chaque enregistrement.

[20] Les données de généalogie pourraient aussi permettre de 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

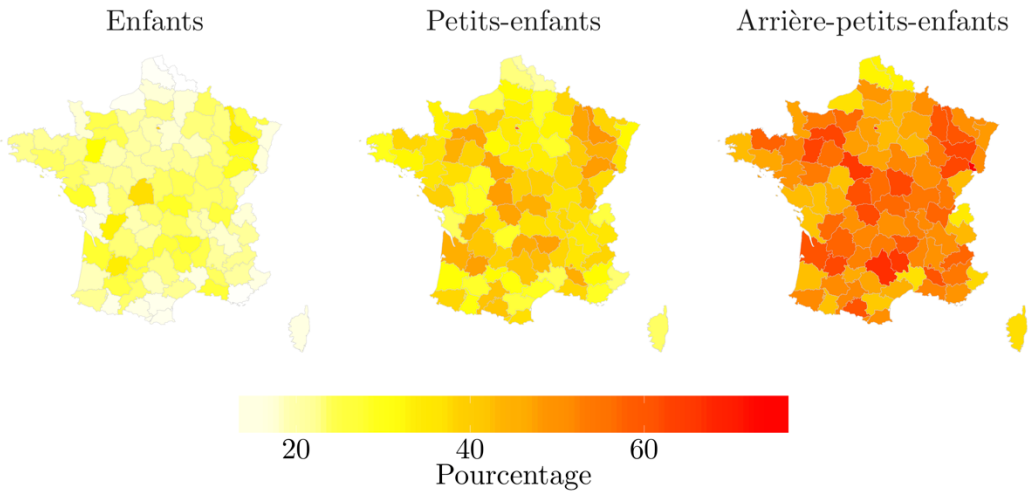


Une première façon de décrire les migrations intergénérationnelles consiste à regarder à l'échelle des départements s'il existe des déplacements conséquents de la population. Pour ce faire, nous comparons le département de naissance de nos individus nés entre 1800 et 1804 en France avec celui de leurs descendants. Dans ce cadre, nous considérons des couples (individu des générations 1800-1804, descendant), les descendants étant soit les enfants, soit les petits-enfants, soit les arrière-petits-enfants. Il faut noter qu'avec ce choix de représentation, un enfant pour lequel nous connaissons les deux parents sera présent deux fois dans les observations : une fois pour caractériser la migration par rapport au lieu de naissance de la mère, une seconde fois pour caractériser la migration par rapport au lieu de naissance du père. 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 né entre 1800-1804. La proportion de petits-enfants nés dans un département différent de celui de leur ancêtre augmente légèrement, et il faut attendre la génération des arrière-petits-enfants pour observer des distinctions régionales marquées.

5.1. Migrations de courte et longue distance

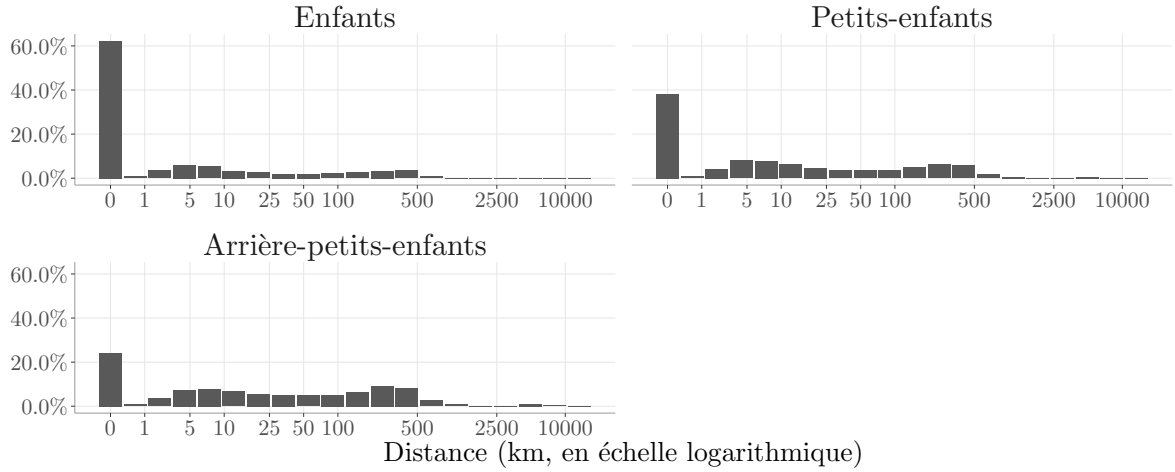
Une analyse plus fine des déplacements est possible en s'intéressant non plus aux départements de naissance, mais aux distances parcourues par les individus d'une génération à la suivante. Plus précisément, en utilisant les coordonnées des lieux de naissance des individus né en France entre 1800 et 1804 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 migration est alors observée d'une manière différente de ce qui est proposé dans la partie précédente : à présent, nous regardons les migrations d'un descendant (enfant, petit-enfant ou arrière-petit-enfant) relativement au lieu de naissance de son aïeul né entre 1800 et 1804 « le plus proche ». Par cette notion de « plus proche aïeul », nous faisons référence à l'aïeul né entre 1800 et 1804 dont le lieu de naissance se situe géographiquement le plus proche de celui du descendant. Ainsi, pour un individu de la 1ère génération, son aïeul né entre 1800 et 1804 le plus proche sera celui parmi ses deux parents dont le lieu de naissance est géographiquement le plus proche de son propre lieu de naissance. Pour un individu de la 2e génération, son aïeul né entre 1800 et 1804 le plus proche sera celui parmi ses quatre grands parents dont le lieu de naissance est géographiquement le plus proche de son propre lieu de naissance, et ainsi de suite. La Figure 8 propose une visualisation graphique de ces distances. À l'instar des résultats mis en avant par Bourdieu, Postel-Vinay, Rosental, & Suwa-Eisenmann (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 sédentaires.

Figure 7. Pourcentage de descendants nés dans un département différent de celui de leur aïeul né entre 1800-1804, 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.

Figure 8. Distribution des distances séparant le lieu de naissance des individus de l'échantillon du lieu de naissance de leur plus proche aïeul né entre 1800 et 1804.



Dans la littérature, on caractérise fréquemment les déplacements des individus en fonction de la distance parcourue, en distinguant 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 distance, 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 le Tableau 2, la part de migrants de proximité est quasiment égale à celle de migrants de longue distance (19% et 18%, 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 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. Il faut toutefois rester prudent sur l'interprétation de ces résultats. En effet, il est possible que les pourcentages élevés de sédentaires soient en partie dus à la relative difficulté pour les généalogistes amateurs de retrouver des ancêtres mobiles.

Tableau 2. Répartition des individus de l'échantillon selon la distance entre leur lieu de naissance et celui de leur plus proche aïeul né entre 1800 et 1804(en pourcentages).

	Enfants	Petits-enfants	Arrière-petits-enfants
Sédentaires	62	38	24
Migrants de proximité	20	28	27
Migrants de longue distance	18	34	49
Total	100	100	100

Note : Les sédentaires sont nés dans le même lieu que leur aïeul le plus proche né entre 1800-1804, les migrants de proximité sont nés à moins de 20 km d'écart et les migrants de longue distance à plus de 20 km.

6. Discussion et conclusion

Dans cet article, nous avons exploité des données issues de la collaboration de centaines de milliers de passionnés de généalogie construisant et partageant leur arbre généalogique. Ces données ont été utilisées pour étudier la population française en métropole, au cours du XIXe siècle. Bien que les sources sur lesquelles s'appuient les généalogistes coïncident avec celles qu'utilisent certains historiens et démographes, à savoir les registres paroissiaux et civils, leur numérisation est totalement différente. En effet, dans le cas des généalogistes, chaque passionné construit son propre arbre. Il peut indiquer ou non, et avec plus ou moins de précision, ce qu'il a pu trouver sur ses ancêtres. La mise en commun des arbres construits par de nombreux utilisateurs laisse penser qu'il est possible d'exploiter la richesse contenue dans chacun des arbres, notamment dans un objectif de description de la population dans le passé. Toutefois, tandis que le travail méthodique et rigoureux fourni par des chercheurs dans le cadre de projets de recherche tels que l'enquête TRA permet de minimiser l'occurrence de biais dans les données, il n'est pas certain que la mise en commun des données de généalogie partagées par des passionnés ne comportent pas de biais. Ainsi, l'objectif principal de cet article est d'explorer les limites des données collaboratives de généalogie. Les objectifs secondaires visent à montrer

d'autres champs d'applications possibles, notamment l'étude de la mortalité et celle de la migration.

Nos résultats montrent qu'il n'est pas possible de s'appuyer sur ces données pour étudier la fécondité des femmes, du fait de la présence de forts biais la sous-estimant. Le manque de report de naissances semble par ailleurs avoir des répercussions sur la mortalité infantile, qu'il convient également d'éviter d'étudier à l'aide des données de généalogie. L'étude de la mortalité n'est pour autant pas impossible, dès lors qu'il s'agit de se concentrer sur une population plus âgée, ayant survécu aux bas-âges. En effet, les résultats obtenus sur l'estimation de la survie des individus ayant dépassé la vingtaine correspondent à ceux avancés dans la littérature. Par ailleurs, l'étude de la migration des individus donne des résultats similaires à ceux que l'on peut trouver dans la littérature, notamment sur les distances parcourues par les individus d'une génération à l'autre.

7. Remerciements

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.

Ces travaux ont également été soutenus par la subvention de l'Agence nationale de la recherche (ANR-17-EURE-0020).

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 avions pu avoir alors, et qui ont motivé certains éléments présentés dans cette étude. Nous remercions également les membres de l'unité « Histoire et Populations » de l'INED pour leurs commentaires. Nous avons bénéficié d'échanges fructueux avec les participants de différentes conférences, parmi lesquelles : « UseR » (Budapest, mai 2018), les « rencontres R » (Rennes, juillet 2018), la «XXIX International Biometric Conference » (Barcelone, juillet 2018), le séminaire « Eco-lunch » (Marseille, septembre 2018), l' « École thématique TEPP-CNRS, Évaluation des Politiques Publiques » (Aussois, mars 2019).

8. Références

Balsamo, G. (1999). *Pruning the genealogical tree: procreation and lineage in literature, law, and religion*. Bucknell University Press.

Bean, L. L., May, D. L., & Skolnick, M. (1978). The Mormon Historical Demography Project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *11*(1), 45–53.

Blayo, Y., & Henry, L. (1967). Données démographiques sur la Bretagne et l'Anjou de 1740 à 1829. *Annales de démographie historique*, *1967*(1), 91–171.

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

Bourdieu, J., Kesztenbaum, L., & Postel-Vinay, G. (2014). L'enquête TRA, une matrice d'histoire. *Population*, *69*(2), 217–248.

Bourdieu, J., Postel-Vinay, G., Rosental, P.-A., & 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. https://doi.org/10.3406/ahess.2000.279879

Bourdieu, J., Postel-Vinay, G., Rosental, P.-A., & 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.

Brian, E., & Jaisson, M. (2007). Sex Ratios at Birth and the Calculus of Probabilities. In *The Descent of Human Sex Ratio at Birth* (p. 221–229). Consulté à l'adresse https://doi.org/10.1007/978-1-4020-6036-6

Brunet, G., & Bideau, A. (2001). Démographie historique et généalogie. *Annales de démographie historique*, *2000*(2), 101–110. https://doi.org/10.3406/adh.2001.1977

Brunet, G., & Vézina, H. (2015). Les approches intergénérationnelles en démographie historique. *Annales de démographie historique*, *129*(1), 77.

Chesnais, J. C. (1986). *La Transition démographique: etapes, formes, implications économiques*. Presses universitaires de France.

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

Dupâquier, J. (1981). Une grande enquête sur la mobilité géographique et sociale aux XIXe et XXe siècles. *Population*, *36*(6), 1164–1167.

Dupâquier, J. (1993). Généalogie et démographie historique. *Annales de démographie historique*, *1993*(1), 391–395. https://doi.org/10.3406/adh.1993.1851

Dupâquier, J., & Blanchet, D. (1992). *La société française au XIXe siècle: tradition, transition, transformations*. Fayard.

Dupâquier, J., & Kessler, D. (1992). L'enquête des 3 000 familles. In J. Dupâquier & D. Kessler (Éd.), *La société française au XIXe siècle: tradition, transition, transformations* (p. 23–61). Fayard.

Eichner, C. J. (2014). In the Name of the Mother: feminist opposition to the patronym in Nineteenth-Century France. *Signs: Journal of Women in Culture and Society*, *39*(3), 659–683.

Fire, M., & 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. https://doi.org/10.1145/2700464

Fleury, M., & Henry, L. (1958). Pour connaitre la population de la France depuis Louis XIV. Plan de travaux par sondage. *Population*, *13*(4), 663. https://doi.org/10.2307/1525088

Gavrilov, L. A., & Gavrilova, N. S. (2001). Etude biodemographique des determinants familiaux de la longevite humaine. *Population*, *56*(1/2), 225.

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

Gavrilova, N. S., & Gavrilov, L. A. (2007). Search for Predictors of Exceptional Human Longevity. *North American Actuarial Journal*, *11*(1), 49–67. https://doi.org/10.1080/10920277.2007.10597437

Girres, J.-F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, *14*(4), 435–459. 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., … 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.

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.

Henry, L., & Blayo, Y. (1975). La population de la France de 1740 a 1860. *Population*, *30*, 71.

Henry, L., & Houdaille, J. (1979). Célibat et age au mariage aux XVIIIe et XIXe siècles en France. II. Age au premier mariage. *Population*, *34*(2), 403. https://doi.org/10.2307/1531570

Hervis, C. (2012). Généalogie : les nouvelles demandes du collectionneur, de l'enquêteur et de l'historien. *La Gazette des archives*, *227*(3), 27–32. https://doi.org/10.3406/gazar.2012.4955

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. https://doi.org/10.2307/1529863

Hollingsworth, T. H. (1976). Genealogy and historical demography. *Annales de démographie historique*, *1976*(1), 167–170. https://doi.org/10.3406/adh.1976.1310

Houdaille, J., & Tugault, Y. (1987). Une bourgeoisie peu malthusienne dans un pays neuf: généalogies américaines du XIXe siècle. *Population (French Edition)*, *42*(2), 305.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., … Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*. 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.

Kesztenbaum, L. (2014). L'étude des migrations grâce aux registres matricules militaires. *Popolazione e storia*, *14*(2), 9–38.

Lease, M., & Yilmaz, E. (2013). Crowdsourcing for information retrieval: introduction to the special issue. *Information Retrieval*, *16*(2), 91–100.

Lindahl-Jacobsen, R., Hanson, H. A., Oksuzyan, A., Mineau, G. P., Christensen, K., & 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.

Matthijs, K., & Moreels, S. (2010). The Antwerp COR*-database: A unique Flemish source for historical-demographic research. *The History of the Family*, *15*(1), 109–115.

Rosental, P.-A. (2004). La migration des femmes (et des hommes) en France au XIXe siècle. *Annales de démographie historique*, *107*(1), 107. 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*.

Toulemon, L. (1995). Très peu de couples restent volontairement sans enfant. *Population*, *50*(4/5), 1079.

Vallin, J., & 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.

van de Walle, E. (1986). La fécondité française au XIXe siècle. *Communications*, *44*(1), 35–45. https://doi.org/10.3406/comm.1986.1653

Warby, S. C., Wendt, S. L., Welinder, P., Munk, E. G. S., Carrillo, O., Sorensen, H. B. D., … Mignot, E. (2014). Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods*, *11*(4), 385–392.

Annexe de l'article « Étude de la démographie française du XIX[e] siècle à partir de données collaboratives de généalogie »

Arthur Charpentier[a,b] & Ewen Gallic[21,c]

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

Mai 2019

1. Intervalles de confiance

1.1. Intervalles de confiance pour le rapport de masculinité

Il est possible de s'appuyer sur les taux d'échantillonnage par département, à savoir le rapport entre la taille de l'échantillon sur la taille de la population recensée par l'Insee, afin de fournir un intervalle de confiance pour la mesure du rapport de masculinité. Le détail du calcul est proposé dans les annexes de Brian & Jaisson . Les valeurs que nous obtenons sont reportées dans le Tableau 3.

Tableau 3. Rapport de masculinité à la naissance par département avec intervalles de confiance, pour les individus nés en 1801.

Département	Rapport de masculinité	Intervalle de confiance à 95%	Département	Rapport de masculinité	Intervalle de confiance à 95%
Ain (01)	110,7	[95,4 ; 128,5]	Lot (46)	121,1	[99,3 ; 148,4]
Aisne (02)	121,1	[108,6 ; 135,3]	Lot-et-Garonne (47)	114,1	[91,7 ; 142,5]
Allier (03)	112,5	[98,6 ; 128,5]	Lozère (48)	120,2	[96,9 ; 149,7]
Alpes-de-Haute-Provence (04)	113,2	[91,9 ; 139,7]	Maine-et-Loire (49)	120,4	[105,3 ; 137,9]
Hautes-Alpes (05)	124,3	[103,7 ; 149,6]	Manche (50)	121,1	[108,6 ; 135,1]
Alpes-Maritimes (06)	107,4	[88,4 ; 130,8]	Marne (51)	123,6	[110,2 ; 138,9]
Ardèche (07)	123,2	[106,8 ; 142,3]	Haute-Marne (52)	120,9	[105,4 ; 139]
Ardennes (08)	121,6	[104,9 ; 141,2]	Mayenne (53)	117,3	[102,5 ; 134,5]
Ariège (09)	112,7	[91,4 ; 139,4]	Meurthe-et-Moselle (54)	126	[112,9 ; 140,9]
Aube (10)	127,1	[106,5 ; 152,3]	Meuse (55)	122,5	[108,6 ; 138,5]
Aude (11)	128,2	[109,3 ; 150,9]	Morbihan (56)	116,6	[105 ; 129,7]
Aveyron (12)	116,3	[100,9 ; 134,3]	Moselle (57)	113,8	[103,7 ; 125]
Bouches-du-Rhône (13)	105,8	[88,7 ; 126,3]	Nièvre (58)	122,7	[106,2 ; 142,1]
Calvados (14)	121	[102,5 ; 143,2]	Nord (59)	117,2	[109,9 ; 124,9]

[21] Auteur correspondant. *Correspondance* : École d'Économie d'Aix-Marseille, Aix-Marseille Université, 5-9 Boulevard Bourdet, CS 50498, 13205 Marseille Cedex 1, France. *Courriel* : ewen.gallic@univ-amu.fr

Cantal (15)	122,1	[102,3 ; 146,2]	Oise (60)	124,4	[109 ; 142,2]
Charente (16)	117,4	[98,3 ; 140,7]	Orne (61)	128	[109,8 ; 149,7]
Charente-Maritime (17)	111,6	[98 ; 127,1]	Pas-de-Calais (62)	114,7	[105,7 ; 124,6]
Cher (18)	109,3	[96,7 ; 123,6]	Puy-de-Dôme (63)	120,9	[106,9 ; 137]
Corrèze (19)	117,8	[101,3 ; 137,3]	Pyrénées-Atlantiques (64)	110,5	[97,7 ; 125,1]
Corse (20)	129,4	[102,7 ; 164]	Hautes-Pyrénées (65)	123,1	[89 ; 172,1]
Côte-D'Or (21)	127,4	[112,7 ; 144,4]	Pyrénées-Orientales (66)	121,6	[99,7 ; 149]
Cotes-D'Armor (22)	123	[112,3 ; 134,9]	Bas-Rhin (67)	109,5	[100,2 ; 119,6]
Creuse (23)	117,7	[101,2 ; 137,2]	Haut-Rhin (68)	109,3	[97,4 ; 122,7]
Dordogne (24)	121,3	[106 ; 139]	Rhône (69)	117,7	[102,1 ; 136]
Doubs (25)	118,9	[100,5 ; 141]	Haute-Saône (70)	124,4	[109 ; 142,3]
Drome (26)	136,3	[116,2 ; 160,6]	Saône-et-Loire (71)	116,6	[105,3 ; 129,3]
Eure (27)	109,7	[95,8 ; 125,7]	Sarthe (72)	108,6	[96,1 ; 122,9]
Eure-et-Loir (28)	119,7	[103,9 ; 138,1]	Savoie (73)	108,5	[94,1 ; 125,2]
Finistère (29)	111,6	[101,3 ; 122,9]	Haute-Savoie (74)	109,5	[94,6 ; 126,9]
Gard (30)	121,7	[100,6 ; 147,8]	Paris (75)	125,7	[103,4 ; 153,5]
Haute-Garonne (31)	110,8	[94,8 ; 129,8]	Seine-Maritime (76)	120,7	[108 ; 135,1]
Gers (32)	125,7	[90,7 ; 176,5]	Seine-et-Marne (77)	123	[109,3 ; 138,6]
Gironde (33)	114,7	[97,7 ; 135]	Seine-et-Oise (78)	118	[105 ; 132,6]
Hérault (34)	111	[94,9 ; 130]	Deux-Sèvres (79)	112,4	[98 ; 129,1]
Ille-et-Vilaine (35)	119,3	[107,6 ; 132,3]	Somme (80)	120,6	[107,5 ; 135,5]
Indre (36)	118,3	[100,9 ; 138,9]	Tarn (81)	108,2	[93,3 ; 125,6]
Indre-et-Loire (37)	127,8	[106,1 ; 154,8]	Tarn-et-Garonne (82)	119,3	[96,4 ; 148,1]
Isère (38)	118,9	[105,7 ; 133,8]	Var (83)	117,8	[95,7 ; 145,6]
Jura (39)	118,9	[99,8 ; 142]	Vaucluse (84)	120	[99,9 ; 144,6]
Landes (40)	114,6	[96,7 ; 136,1]	Vendée (85)	116,1	[103,1 ; 130,9]
Loir-et-Cher (41)	112,7	[95,8 ; 132,9]	Vienne (86)	112	[97,2 ; 129,4]
Loire (42)	105,1	[90,8 ; 121,7]	Haute-Vienne (87)	128,9	[108,5 ; 153,7]
Haute-Loire (43)	117,1	[99,6 ; 138]	Vosges (88)	125,2	[111,9 ; 140,3]
Loire-Atlantique (44)	124,8	[111,3 ; 140,2]	Yonne (89)	117,7	[104 ; 133,3]
Loiret (45)	116,3	[100,9 ; 134,2]	Territoire de Belfort (90)	123,3	[89,3 ; 172,2]

1.2. Intervalles de confiance par bootstrap pour l'espérance de vie

Les estimations de d'espérance de vie à la naissance que nous proposons sont accompagnées d'intervalles de confiance à 95% obtenus par *bootstrap*. La méthode que nous utilisons est fortement inspirée de celle expliquée par Carlo Giovanni Camarda sur son site web[22]. Il s'agit, à partir d'une population E_0 née une année donnée (*e.g.*, 1800) et d'une probabilité de décéder dans l'année ($\widehat{q_x}$, estimée sur nos données pour cette cohorte) de procéder de manière itérative. À l'âge x, s'il reste E_x personnes en vie, le nombre de décès D_x est tiré suivant une loi binomiale $\mathcal{B}(E_x, \widehat{q_x})$. Ensuite, nous posons $E_{x+1} = E_x - D_x$,. Nous tirons alors un nouveau nombre de décès pour l'âge suivant, et ainsi de suite. Ces étapes itératives sont reproduites de manière indépendante dans 1 000 échantillons. Dans chacun de ces échantillons, nous calculons

[22] https://sites.google.com/site/carlogiovannicamarda/r-stuff/life-expectancy-confidence-interval.

l'espérance de vie à la naissance selon les tirages effectués. Les quantiles empiriques d'ordres 0,025 et 0,975 deviennent alors les bornes de l'intervalle de confiance à 95% pour l'espérance de vie à la naissance des individus de la cohorte étudiée.

Les données de généalogie souffrent d'un biais lorsqu'il s'agit d'estimer les espérances de vie à la naissance. Toutefois, il est intéressant de se pencher sur l'hétérogénéité spatiale des estimations. Pour ce faire, nous estimons l'espérance de vie à la naissance dans chacun des départements, pour les cohortes d'individus de 1800 à 1804.

La Figure A.1 propose une représentation graphique de l'hétérogénéité spatiale des estimations d'espérance de vie à la naissance par département, en reportant la longueur de l'intervalle de confiance associé à chaque mesure. Les différences régionales qui se dégagent sont indépendantes du sexe des individus. Dans les régions correspondant à l'actuelle Occitanie ainsi que la Normandie, l'espérance de vie des individus nés entre 1800 et 1804 est relativement plus élevée que dans les autres régions. En revanche, le centre de la France semble touché par une faible espérance de vie des individus, relativement au reste de la France.

Par ailleurs, la Figure A.2 propose de mettre en lien nos estimations départementales d'espérance de vie à la naissance (points et barres rouges pour les femmes, bleus pour les hommes) avec les valeurs avancées par van de Walle (1973) (triangles noirs).

Figure 1.1. Hétérogénéité spatiale des espérances de vie à la naissance par département pour les cohortes de 1800 à 1804.

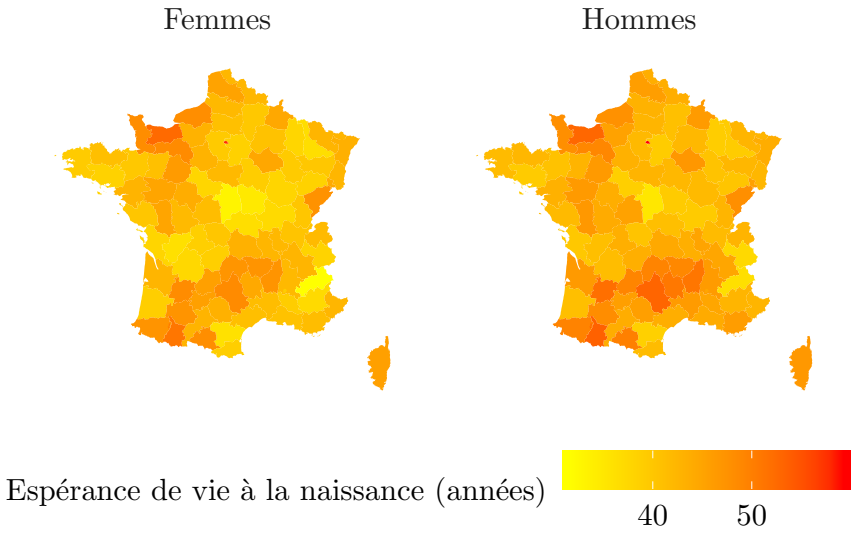
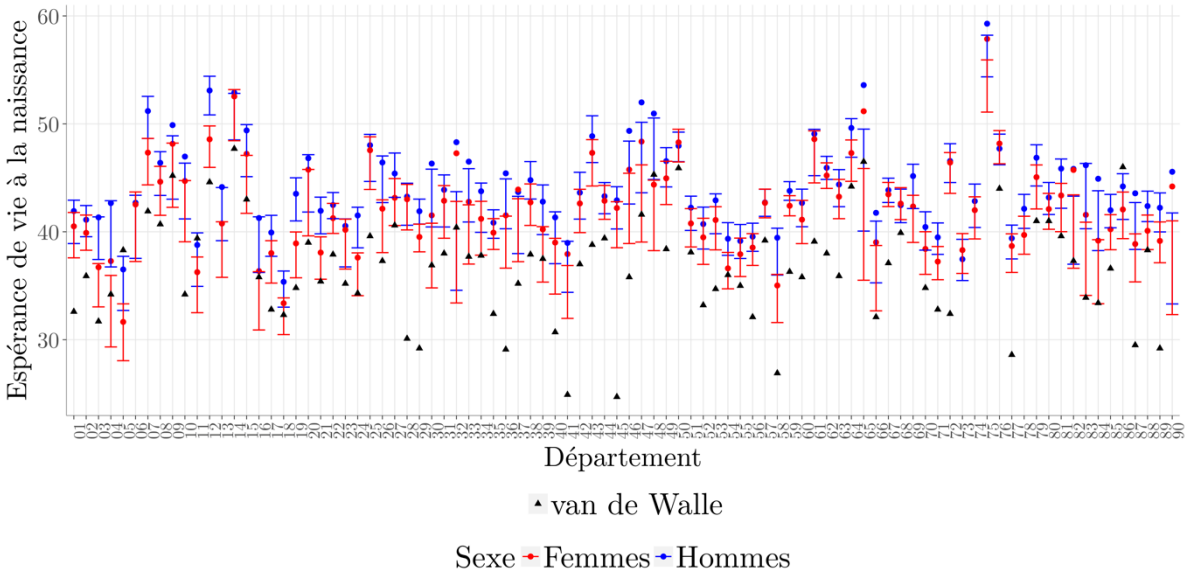


Figure 1.2. Espérances de vie à la naissance, par département.



Note : Les estimations de l'espérance de vie à la naissance par département sont représentées par des points (rouges pour les femmes, bleus pour les femmes). Les estimations ponctuelles proposées par van de Walle (1973) sont représentées par les triangles noirs. 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.

Tableau 4. Espérance de vie à la naissance par département et par genre pour les cohortes de 1800 à 1804, avec intervalles de confiance à 95%.

	Femmes		Hommes			Femmes		Hommes	
Département	**Espérance de vie**	**I.C. 95%**	**Espérance de vie**	**I.C. 95%**	**Département**	**Espérance de vie**	**I.C. 95%**	**Espérance de vie**	**I.C. 95%**
Ain (01)	39,7	[37,6 ; 41,8]	40,9	[38,9 ; 42,9]	Indre-et-Loire (37)	40,1	[37,2 ; 43,1]	40,6	[38 ; 43,3]
Aisne (02)	39,9	[38,3 ; 41,5]	41	[39,5 ; 42,4]	Isère (38)	42,5	[40,6 ; 44,4]	44,8	[43 ; 46,5]
Allier (03)	35	[33 ; 37,1]	39,4	[37,4 ; 41,4]	Jura (39)	37,9	[35,3 ; 40,5]	42	[39,7 ; 44,3]
Alpes-de-Haute-Provence (04)	32,6	[29,3 ; 36]	39,8	[36,7 ; 42,9]	Landes (40)	36,8	[34,2 ; 39,4]	39,4	[37 ; 41,8]
Alpes-Maritimes (06)	40,5	[37,2 ; 43,7]	40,4	[37,5 ; 43,4]	Loir-et-Cher (41)	34,5	[32 ; 36,9]	36,7	[34,4 ; 39,1]
Ardèche (07)	46,5	[44,3 ; 48,7]	50,6	[48,6 ; 52,5]	Loire (42)	41,9	[39,9 ; 43,9]	43,3	[41,2 ; 45,5]
Ardennes (08)	43,8	[41,5 ; 46,1]	45,4	[43,4 ; 47,4]	Loire-Atlantique (44)	42,7	[41,2 ; 44,3]	43,1	[41,7 ; 44,6]
Ariège (09)	45,3	[42,3 ; 48,2]	45,9	[43 ; 48,9]	Loiret (45)	40,7	[38,5 ; 42,8]	42,2	[40,3 ; 44,2]
Aube (10)	41,9	[39,1 ; 44,7]	43,7	[41,2 ; 46,4]	Lot (46)	42,2	[38,9 ; 45,4]	45,4	[42,6 ; 48,4]
Aude (11)	35,1	[32,5 ; 37,7]	37,4	[34,9 ; 39,9]	Lot-et-Garonne (47)	42,6	[39 ; 46,2]	46,8	[43,6 ; 50,1]
Aveyron (12)	47,9	[46 ; 49,8]	52,6	[50,8 ; 54,4]	Lozère (48)	41,6	[38,3 ; 44,8]	47,7	[44,8 ; 50,5]
Bas-Rhin (67)	43,4	[42,3 ; 44,6]	43,8	[42,7 ; 45]	Maine-et-Loire (49)	44,5	[42,5 ; 46,5]	45,9	[44,2 ; 47,8]
Bouches-du-Rhône (13)	38,3	[35,8 ; 40,9]	41,6	[39,2 ; 44,1]	Manche (50)	48	[46,5 ; 49,5]	47,9	[46,5 ; 49,2]
Calvados (14)	50,8	[48,4 ; 53,2]	50,6	[48,5 ; 52,8]	Marne (51)	40,3	[38,6 ; 42,1]	41,7	[40,1 ; 43,3]
Cantal (15)	44,4	[41,7 ; 47,1]	47,5	[45,1 ; 49,9]	Mayenne (53)	40,3	[38,3 ; 42,3]	41,6	[39,8 ; 43,5]
Charente (16)	33,6	[30,9 ; 36,3]	38,8	[36,2 ; 41,4]	Meurthe-et-Moselle (54)	36,4	[34,7 ; 38,1]	39,3	[37,8 ; 40,8]
Charente-Maritime (17)	37,2	[35,2 ; 39,2]	39,6	[37,7 ; 41,5]	Meuse (55)	37,6	[35,9 ; 39,4]	39,1	[37,5 ; 40,7]

Cher (18)	32,1	[30,5 ; 33,9]	34,7	[33 ; 36,4]	Morbihan (56)	38,4	[36,9 ; 39,8]	39,5	[38,2 ; 40,8]
Corrèze (19)	37,9	[35,7 ; 40]	43	[41 ; 45]	Moselle (57)	42,6	[41,3 ; 44]	42,7	[41,4 ; 43,9]
Corse (20)	42,7	[39,6 ; 45,8]	44,5	[41,8 ; 47,1]	Nièvre (58)	33,8	[31,6 ; 36]	38,1	[36 ; 40,3]
Côte-D'Or (21)	37,6	[35,6 ; 39,5]	41,6	[39,8 ; 43,2]	Nord (59)	42,4	[41,5 ; 43,3]	43,8	[42,9 ; 44,6]
Cotes-D'Armor (22)	41,2	[39,8 ; 42,6]	42,5	[41,2 ; 43,6]	Oise (60)	40,9	[38,9 ; 42,9]	42,2	[40,5 ; 44]
Creuse (23)	38,8	[36,5 ; 41,2]	38,9	[36,7 ; 41,2]	Orne (61)	46,9	[44,5 ; 49,4]	47,3	[45,2 ; 49,5]
Deux-Sèvres (79)	44,2	[42,1 ; 46,2]	46,1	[44,2 ; 48,1]	Paris (75)	53,6	[51,1 ; 55,9]	56,3	[54,4 ; 58,2]
Dordogne (24)	36,1	[34,1 ; 38]	40,4	[38,5 ; 42,3]	Pas-de-Calais (62)	45,2	[44 ; 46,4]	45,9	[44,9 ; 47]
Doubs (25)	46,4	[43,9 ; 48,8]	46,8	[44,7 ; 49]	Puy-de-Dôme (63)	43	[41,2 ; 44,9]	44	[42,3 ; 45,8]
Drome (26)	40,5	[38,1 ; 43]	44,9	[42,7 ; 47]	Pyrénées-Atlantiques (64)	46,6	[44,7 ; 48,5]	48,7	[46,9 ; 50,5]
Eure (27)	42,8	[40,6 ; 44,9]	45,2	[43,2 ; 47,3]	Pyrénées-Orientales (66)	35,7	[32,7 ; 38,7]	38,1	[35,3 ; 41]
Eure-et-Loir (28)	42,3	[40,2 ; 44,4]	42,6	[40,6 ; 44,5]	Rhône (69)	41,2	[39 ; 43,4]	44,2	[42,2 ; 46,3]
Finistère (29)	39,4	[38,1 ; 40,7]	41,9	[40,7 ; 43]	Saône-et-Loire (71)	37,1	[35,6 ; 38,6]	39,4	[37,9 ; 40,8]
Gard (30)	37,9	[34,8 ; 40,8]	43	[40,4 ; 45,8]	Sarthe (72)	45,5	[43,6 ; 47,3]	46,3	[44,6 ; 48,2]
Gers (32)	38,1	[33,4 ; 42,8]	39,1	[34,6 ; 43,7]	Savoie (73)	38	[36,1 ; 39,8]	37,4	[35,5 ; 39,3]
Gironde (33)	39,8	[37 ; 42,5]	43,4	[40,9 ; 45,8]	Seine-et-Marne (77)	38	[36,2 ; 39,8]	39	[37,5 ; 40,6]
Haut-Rhin (68)	42,6	[41,1 ; 44,1]	42,4	[40,8 ; 44]	Seine-et-Oise (78)	39,7	[37,9 ; 41,4]	41,9	[40,3 ; 43,5]
Haute-Garonne (31)	41,9	[39,4 ; 44,3]	42,8	[40,4 ; 45,3]	Seine-Maritime (76)	47,9	[46,3 ; 49,4]	47,7	[46,2 ; 49,1]
Haute-Loire (43)	46,4	[44,2 ; 48,5]	48,5	[46,4 ; 50,7]	Somme (80)	41,9	[40,2 ; 43,6]	43,1	[41,6 ; 44,6]
Haute-Marne (52)	39,1	[37 ; 41,3]	40,4	[38,4 ; 42,3]	Tarn (81)	42,3	[40 ; 44,5]	44,5	[42,2 ; 46,7]
Haute-Saône (70)	38	[36,1 ; 40]	40,1	[38,3 ; 41,9]	Tarn-et-Garonne (82)	40	[36,6 ; 43,4]	40,1	[37 ; 43,3]
Haute-Savoie (74)	41,3	[39,3 ; 43,2]	42,4	[40,4 ; 44,4]	Territoire de Belfort (90)	36,7	[32,3 ; 41]	37,5	[33,3 ; 41,7]
Haute-Vienne (87)	37,5	[35,4 ; 39,8]	40,5	[38,4 ; 42,7]	Var (83)	37,5	[34,1 ; 40,9]	43,3	[40,3 ; 46,3]
Hautes-Alpes (05)	30,6	[28,1 ; 33,3]	35,2	[32,7 ; 37,7]	Vaucluse (84)	36,3	[33,3 ; 39,3]	41	[38,3 ; 43,8]
Hautes-Pyrénées (65)	40,7	[35,5 ; 45,9]	44,8	[40,1 ; 49,5]	Vendée (85)	39,9	[38,3 ; 41,6]	41,9	[40,4 ; 43,5]
Hérault (34)	40,4	[37,8 ; 42,8]	42,2	[39,9 ; 44,5]	Vienne (86)	41,5	[39,4 ; 43,7]	43,3	[41,1 ; 45,4]
Ille-et-Vilaine (35)	39,8	[38,4 ; 41,2]	40,8	[39,4 ; 42]	Vosges (88)	39,9	[38,4 ; 41,5]	42,4	[40,9 ; 43,8]
Indre (36)	39	[36,6 ; 41,5]	42,6	[40,3 ; 44,9]	Yonne (89)	39	[37,1 ; 40,9]	41,8	[40 ; 43,6]