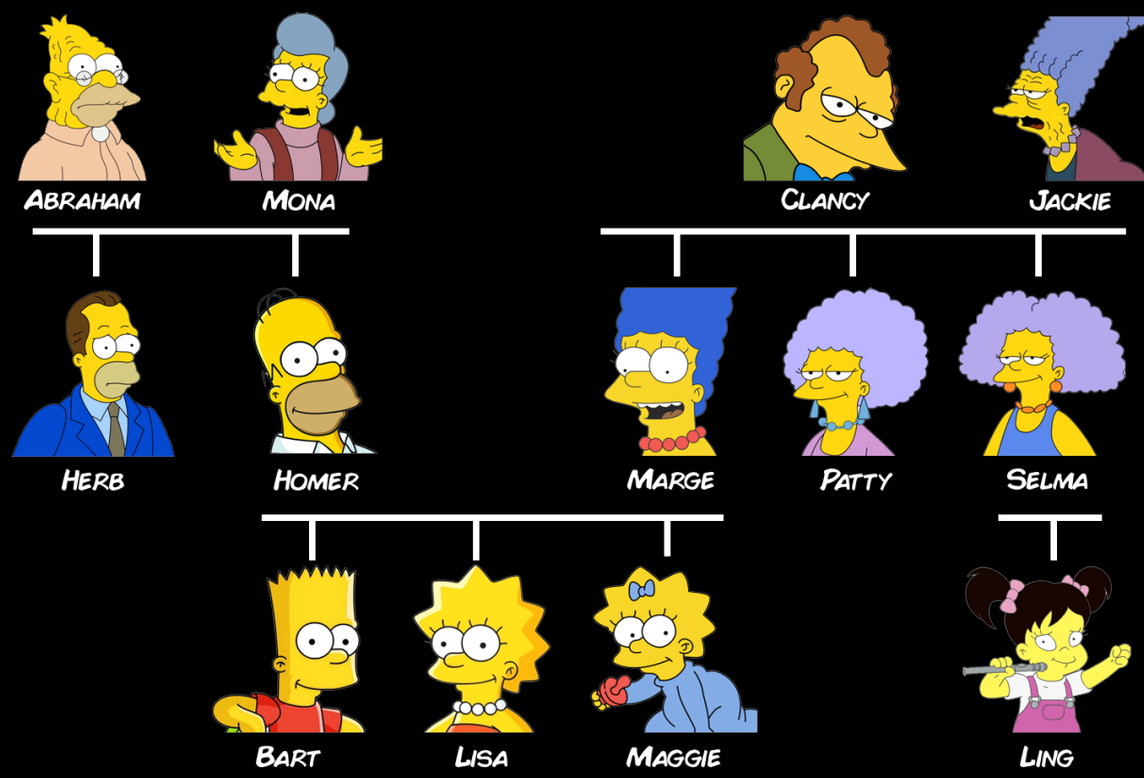
Etude de la démographie française du XIXe siècle à partir de données collaboratives de généalogie

Ewen Gallic & Arthur Charpentier
CREM UMR CNRS 6211, Université de Rennes 1 & Chaire Actinfo



https://3wen.github.io/genealogie_fr

7e rencontres R

Rennes, 4–6 juillet 2018

1. Introduction

Démographie historique : large littérature

- Travaux pionniers de Henry (1956)
- Les **données longitudinales** sont exploitées dans de nombreux projets :
  - Matthijs and Moreels (2010) (COR*)
    - Antwerp, Belgique, 1846–1920, ~125k événements, ~57k individus
  - Mandemakers (2000)
    - Pays-bas, 1812–1922, ~77k individus
  - Bouchard et al. (1989) (BALSAC)
    - Québec, Canada, depuis le 17e siècle, ~2M événements, ~575k individus
  - Bean et al. (1978)
    - principalement Utah, USA, depuis le 18e siècle, ~1.2M individus

1. Introduction

Big data et données collaboratives

- **Peut-on utiliser ces données en démographie historique ?**
- A priori oui, selon les études menées principalement sur la **longévité** :
  - Fire and Elovici (2015) avec des données de WikiTree.com
    - +1M profiles, nombre d'individus non divulgué
  - Cummins (2017) avec des arbres gén. de FamilySearch.org
    - +1,3M d'individus
  - Gavrilova and Gavrilov (2007) avec des données de généalogie de Rootsweb
    - +75M d'individus décédés
  - Gergaud et al. (2016) avec des biographies Wikipédia
    - +1,2M d'individus
  - Kaplanis et al. (2018) avec des arbres généalogiques de Geni.com
    - 13M d'individus
- Ces études ne s'attardent pas sur la **représentativité** de leurs données

2. Méthodologie

Données collaboratives de Geneanet

- Les utilisateurs construisent leur **arbre généalogique**
- Pour chacun des événements (naissance, mariage, décès), ils peuvent indiquer :
  - **des noms**
  - **des dates**
  - **des lieux**

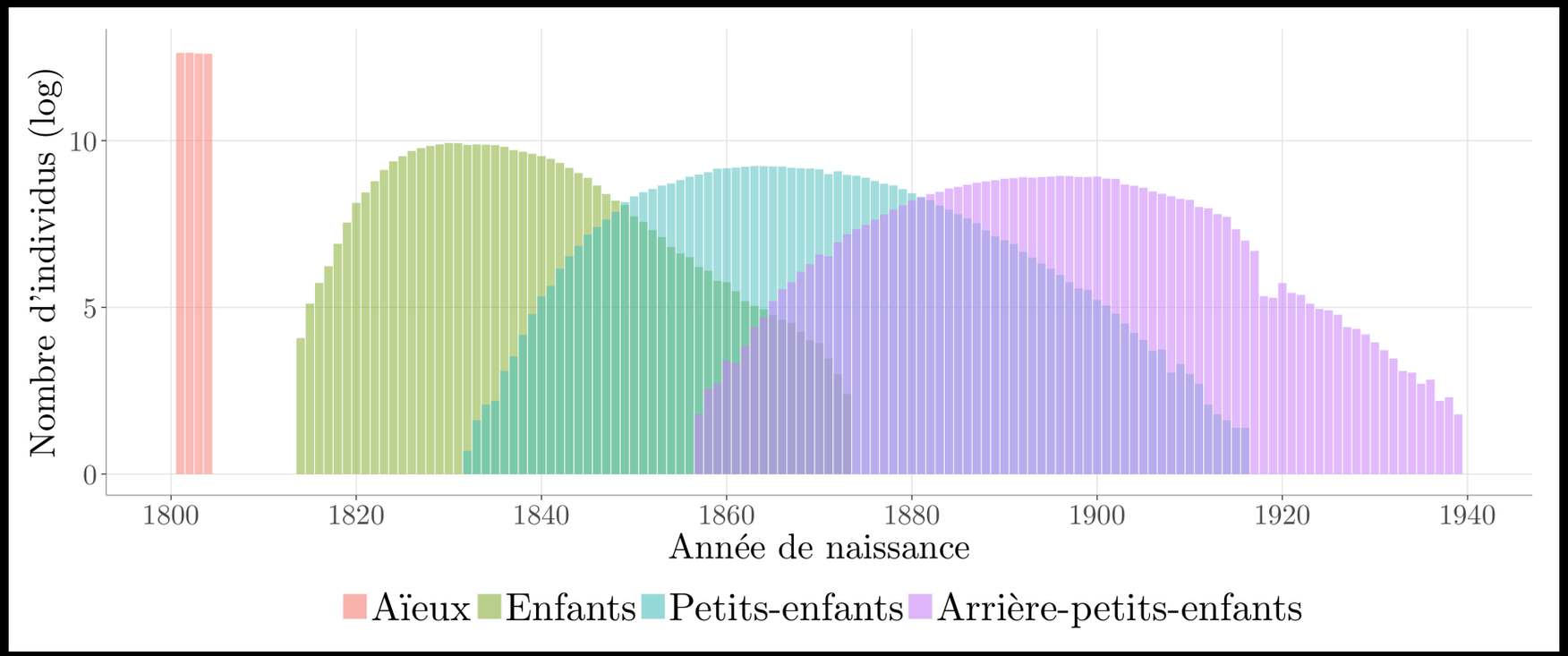
- Extraction à notre disposition :
  - Arbres de 238 009 utilisateurs :
    - **+700M d'enregistrements**
  - parmi ces enregistrements : focus sur les individus nés entre 1800 et 1804 et sur leurs descendants
  - nettoyage de la base

2. Méthodologie

Données collaboratives de Geneanet

- L'échantillon final contient :
  - les individus nés en France (métropolitaine) entre 1800 et 1804
    - 1 547 086 individus
  - leurs descendants jusqu'à 3 générations
    - 403 190 enfants, 286 071 petits-enfants, 222 103 arrière-petits-enfants
- Note : nous n'avons accès qu'aux enregistrements des utilisateurs n'ayant pas refusé de publier leur arbre

Distribution des années de naissance de l'échantillon, par génération.



2. Méthodologie

Construction des arbres à partir des données brutes

Exemple de données brutes.

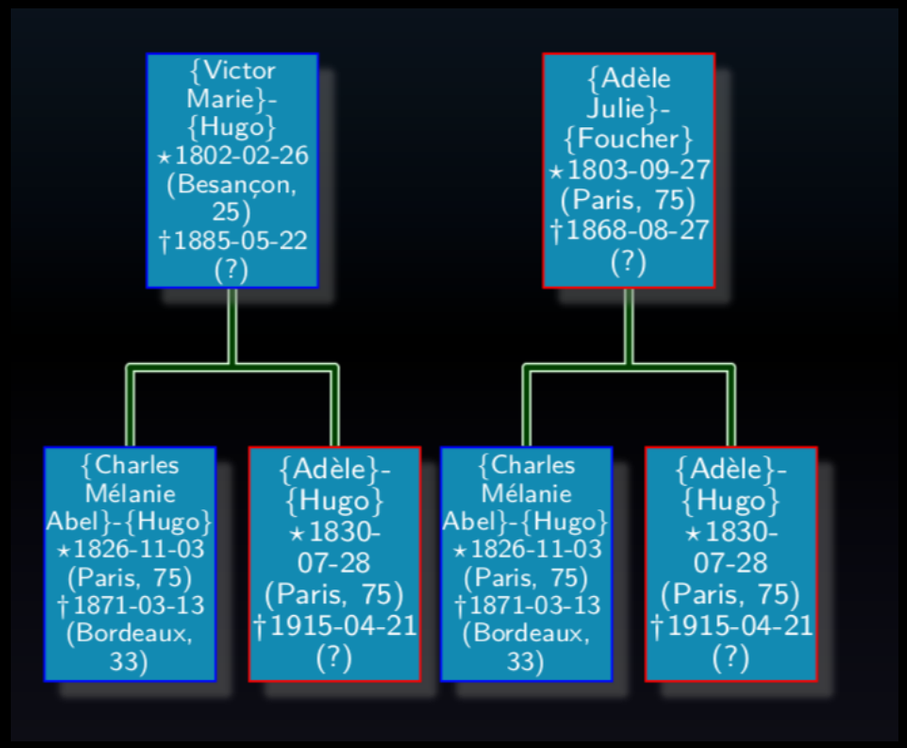
	ID_user	ID_np	ID_num	Name tabular	Surname	Sex	Date_b
1	daage	besnard|jean|1	575	BESNARD	Jean	1	18000227
2	denisgallienne	besnard|louis|1	22771	BESNARD	Louis	1	18040603
3	domiassi	besnard|jean|	1748	BESNARD	Jean	1	18000227
4	dutheilfr	besnard|pierre|	729	BESNARD	Pierre	1	18001221
5	dvivier1	besnard|louis|1	65196	BESNARD	Louis	1	18001215

	Date_d	Type	Location	Lat	Long	ID_num_m	ID_num_p
1	16810000	NM	Longué, 0180	47.37806	-0.10806	4457	574
2	18831027	ND	Cunault, 49350	47.30833	-0.15389	994	1620
3	18560000	NM	Longué, 49180	47.37806	-0.10806		
4		N	Gennes, 49350	47.34083	-0.23278	99	59
5	18490717	N	Pommeraye, 49244	47.35528	-0.86028	43116	4063

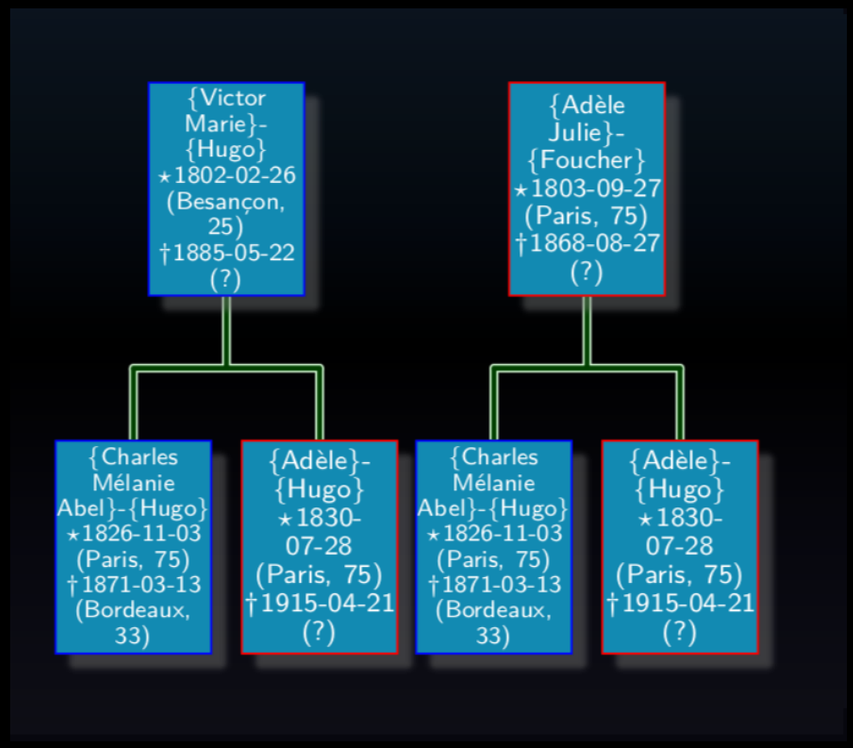
2. Méthodologie

Construction des arbres généalogiques

- Les données brutes correspondent à une **généalogie ascendante**
  - le généalogiste part d'un individu et recherche progressivement les aïeux
- Nous souhaitons suivre les descendants d'individus ; il nous faut donc adopter une **démarche descendante**



(a) Données types de la base : deux individus avec les mêmes parents.



(b) Ce que l'on souhaite : les descendants de chaque parent.

2. Méthodologie

Construction des arbres généalogiques : individus

- Les individus peuvent apparaître plusieurs fois dans les données : **problèmes de doublons**
  - nous les rassemblons à l'aide d'un **algorithme simple**, en six étapes
    - les noms proches (e.g., Jean ou Jehan) sont appariés (lorsque nécessaire) à l'aide d'une mesure de distance entre les chaînes de caractères
- Nous obtenons **2 457 450 individus uniques** dans l'échantillon

Exemple d'individus présents plusieurs fois dans les données.

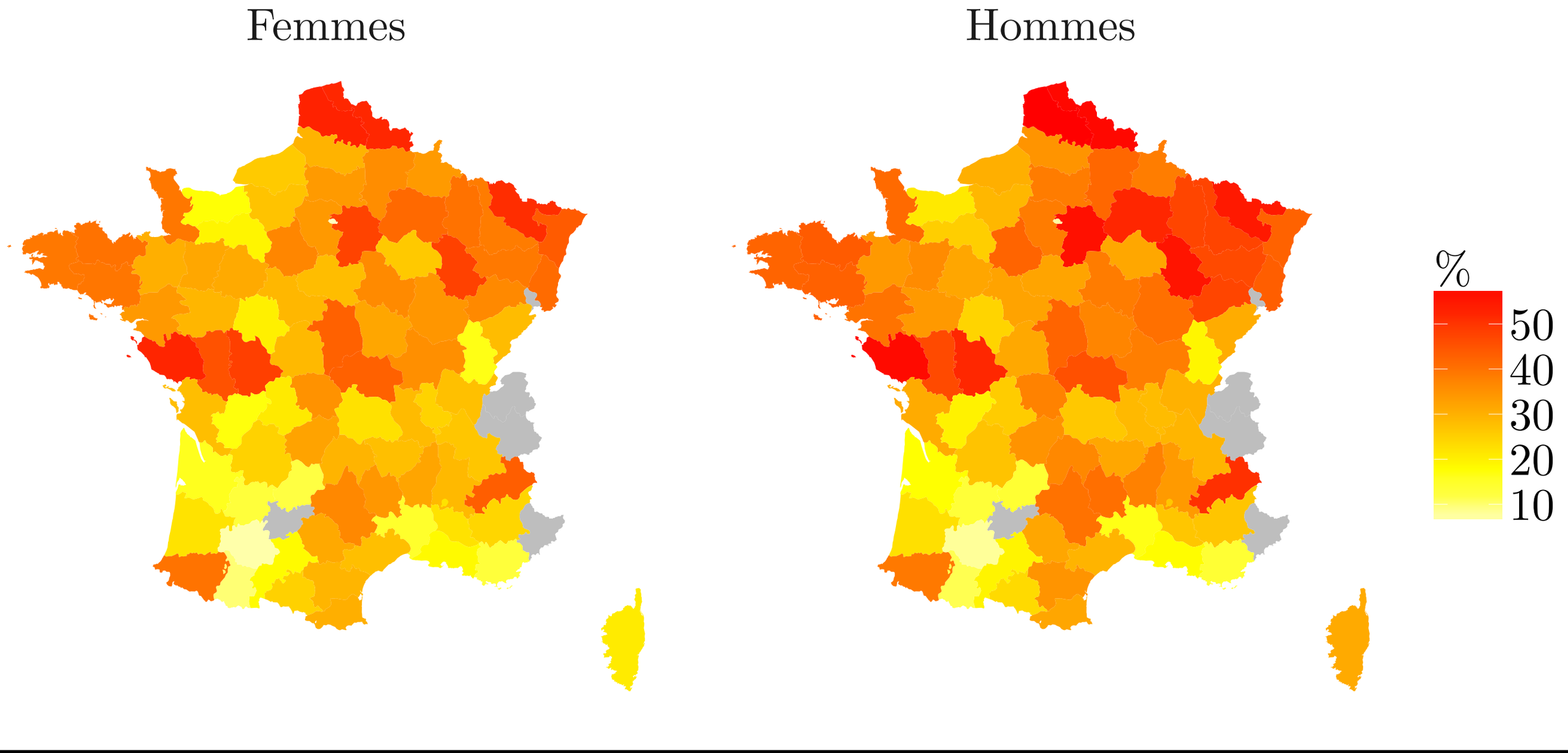
	ID_user	ID_np	ID_num	Nom	Prenom	Sexe	Date_N	Date_D
1	ericde78	jolly|pierre|5	9549	JOLLY	Pierre	1	18000406	
2	cfph89villy	jolly|pierre|8	5142	JOLLY	Pierre	1	18000406	18640120
3	ericde78	fournier|marie brigitte|	6688	FOURNIER	Marie Brigitte	2	17600000	
4	ericde78	jolly|claude|	9487	JOLLY	Claude	1	17570524	18241209
5	cfph89villy	fournier|marie brigitte|	1351	FOURNIER	Marie Brigitte	2	17631005	18271227
6	cfph89villy	jolly|claude|2	1745	JOLLY	Claude	1	17570524	18241209

	Type	Lieu	Lat	Long	ID_num_m	ID_num_p
1	N	Villy,89800	3.75111	47.86778	6688	9487
2	ND	Villy	3.75111	47.86778	1351	1745
3	M	Lignorelles,89800	3.72750	47.86306	495	6713
4	M	Lignorelles,89800	3.72750	47.86306	16871	9547
5	NM	Lignorelles	3.72750	47.86306	167	1360
6	M	Lignorelles	3.72750	47.86306	4236	1906

3. Représentativité

Nombre de naissance : comparaison avec les statistiques officielles

Proportion des naissances par département dans l'échantillon comparativement aux données communiquées par l'INSEE.

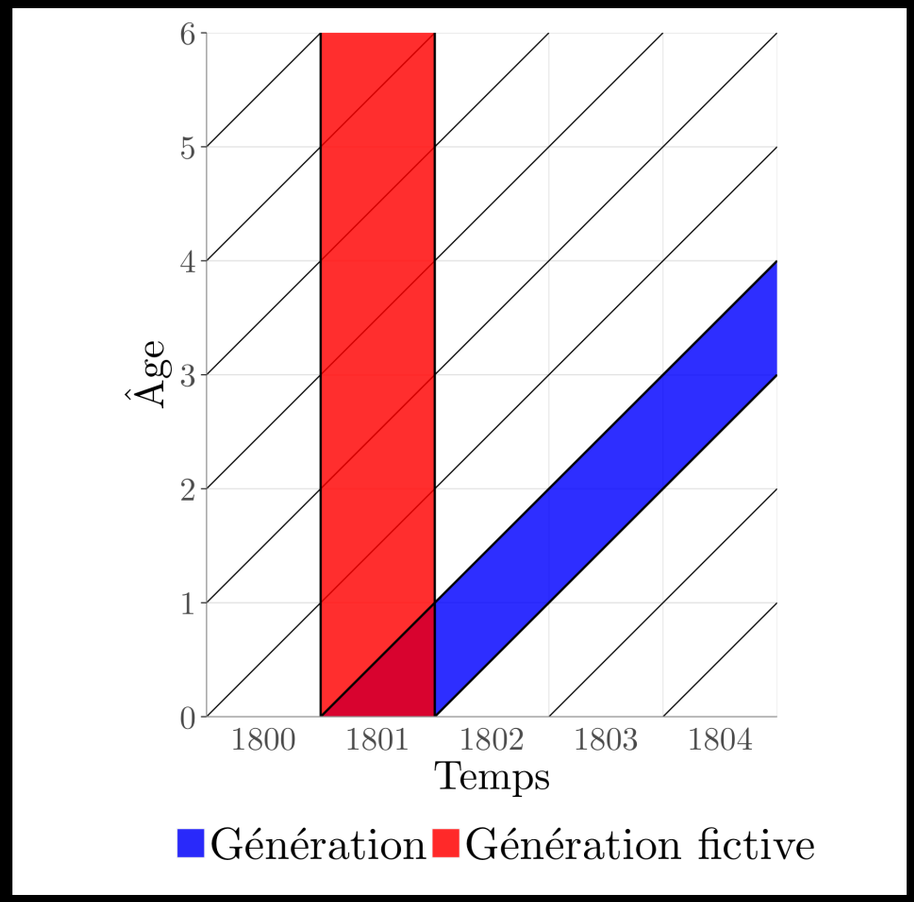


Note : Ces cartes montrent la représentativité de l'échantillon par département en comparant le nombre d'individus nés en 1801 enregistrés dans la base aux enregistrements de l'INSEE. La couleur grise indique une valeur manquante.

4. Mortalité

Suivi par cohorte

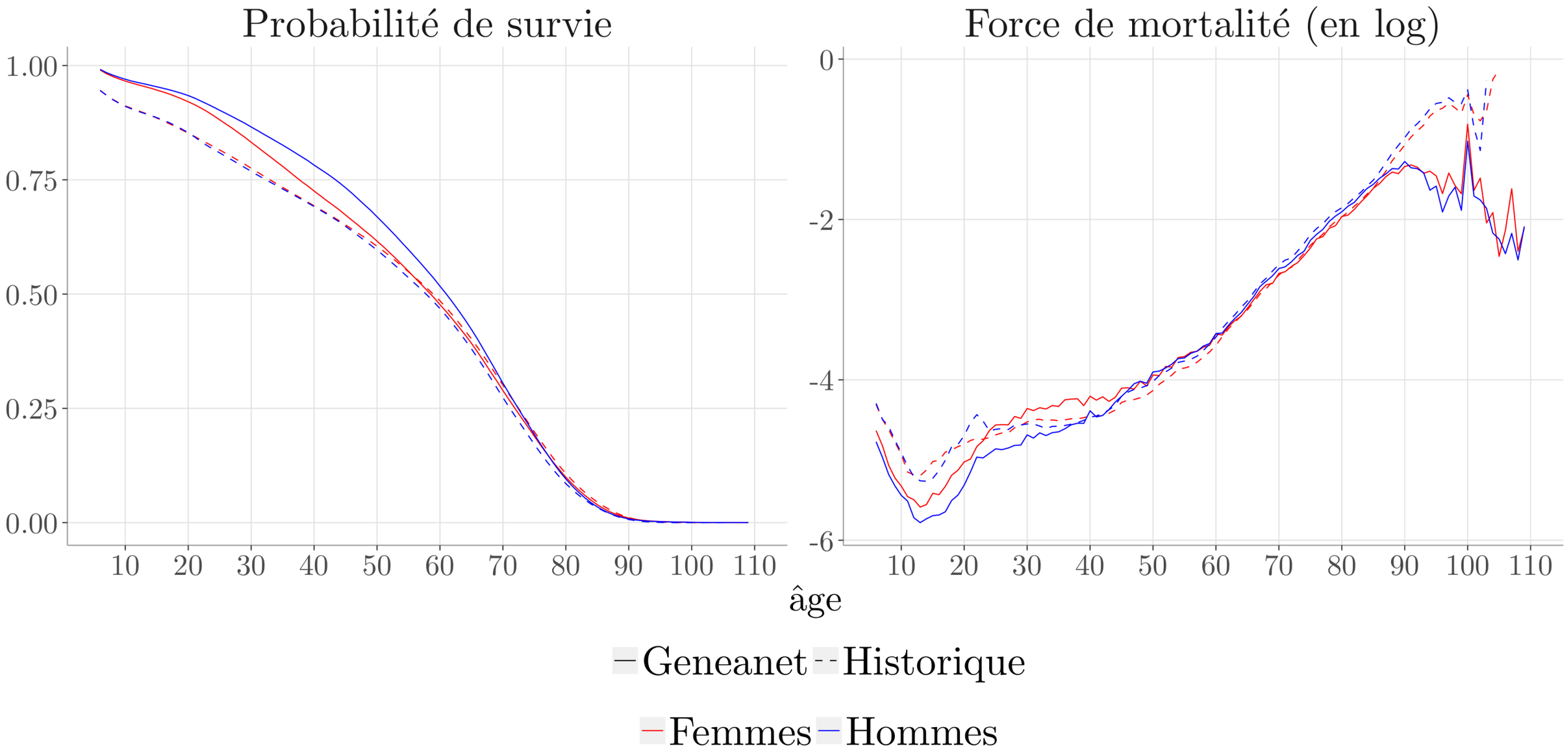
- Nous étudions 5 **cohortes** : aïeux nés entre 1800 et 1804 (813 551 individus)
- Pour chaque âge :
  - combien sont encore en vie ?
  - combien décèdent ?
- Comparaison avec des **tables de mortalité** (Vallin et Meslé, 2001)



4. Mortalité

Analyse de la survie

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.



5. Migration

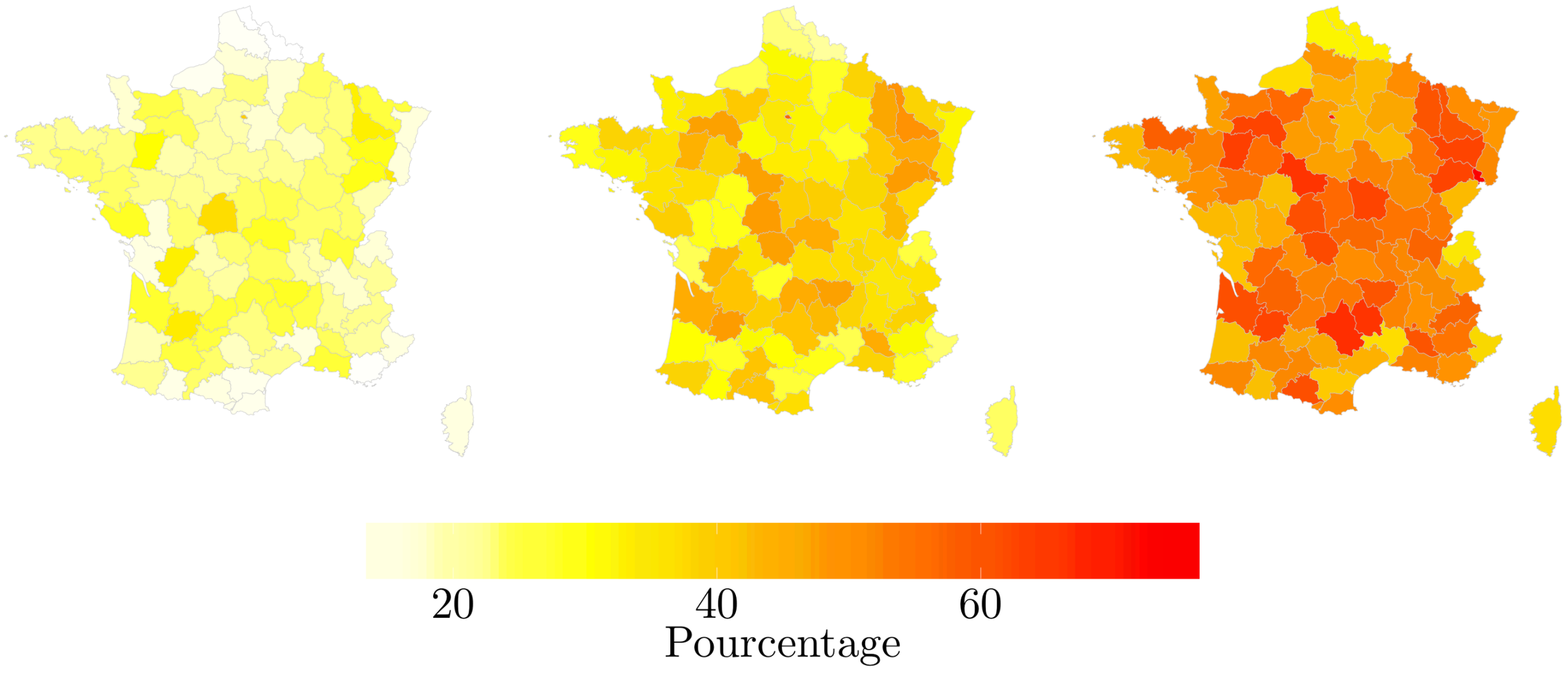
Migration entre les générations

- Exploitation des **coordonnées géographiques** des lieux de naissance
- Migration étudiée à plusieurs échelles spatiales :
  - départements (✓ dans cet exposé)
  - communes (✗ dans le papier uniquement)
  - cas particulier de Paris (✗ dans le papier uniquement)
- Pour le regard à l'échelle des départements :
  - Comparaison entre :
    - le lieu de naissance d'un individu
    - celui de ses descendants
  - Calcul de la **proportion de migrants** (*i.e.*, ici, de descendants à naître dans un département différent que l'aïeul), pour chaque département

5. Migration

Migration entre les générations : à l'échelle des départements

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



Pour aller plus loin...

- **Document de travail**
  - Charpentier, A. and Gallic, E. (2018). Étude de la démographie française du XIXe siècle à partir de données collaboratives de généalogie. hal-01724269
- **Annexe en ligne (codes R)**

https://3wen.github.io/genealogie_fr

Arthur Charpentier (@freakonometrics)

Ewen Gallic (@3wen)

7e rencontres R

Rennes, 4–6 juillet 2018