A Cross-verified Database of Notable People, 3500BC-2018AD

We add to the literature on notable individuals by collecting first a massive amount of data from various editions of Wikipedia and Wikidata along with deduplication techniques; and then using these partially overlapping sources to cross-verify each retrieved information. This strategy results in a cross-verified database of 2.3 million individuals, including a third who are not present in the English edition of Wikipedia. We adopt a social science approach: data collection is driven by specific social questions on gender, economic and cultural development and quantitative exploration of cultural trends.

Project team:

Palaash BARGHAVA

Palaash Bargava is a graduate student in the department of Economics at Columbia University. His research interests are geared towards exploring the influence of networks, culture and non-standard economic preferences on individual labour outcomes and aggregate institutions.

Jean-Benoît EYMÉOUD

Jean-Benoît Eyméoud is an Economist at Banque de France and fellow at LIEPP. He received his PhD in 2018 from Sciences Po under the supervision of Etienne Wasmer. His research lies in urban and political economics with a focus on housing, labor and discrimination issues.

Olivier GERGAUD

Olivier GERGAUD is a senior professor of Economics at KEDGE Business School. His main research interests are Economics of Pro-social Behavior, Cultural Economics Celebrities), Restaurant and Wine (Economics, Environmental Economics, Behavioral Finance (Hedge Funds, Betting) and Sports Economics (Cycling, Football).

Morgane LAOUÉNAN

Morgane Laouénan is a CNRS researcher at the Centre d'Economie de la Sorbonne and co-director of the LIEPP discrimination and category-based policies research group. She is specialized in Labor Economics and Applied Microeconomics. Her research focuses on discrimination in labor markets.

Guillaume PLIQUE

Guillaume Plique joined the Medialab (Sciences Po) in 2013 as Research Engineer. He assists social sciences researchers with the lab's various projects and help them regarding methodology.

Etienne WASMER

Etienne Wasmer is Professor of Economics at New York University Abu Dhabi (NYUAB). He was de co-director of LIEPP between 2011 and 2017. His main research interests lie in labour economics, macroeconomics, search theory and urban economics.

Context

A new strand of literature has emerged that attempts to build the most comprehensive and accurate database of notable individuals. Two recent projects by Schich et al. (2014) and Yu et al., (2016) particularly influenced our work.

We collect a massive amount of data from various editions of Wikipedia and Wikidata. Using deduplication techniques, we cross-verify each retrieved information. For some variables, Wikipedia adds 15% more information when missing in Wikidata. We find very few errors in the part of the database that contains the most documented individuals but nontrivial error rates in the bottom of the notability distribution, due to sparse information and classification errors or ambiguity.

Data collection

We consider two main sources of information: Wikidata and Wikipedia (7 language editions).

Wikipedia: we analyze the source code of each Wikipedia biography to extract the following information: birth and death (date and location), main occupations, gender and citizenship.

Wikidata: we use the "instance of humans" category to figure out a sample of individuals in this universe. We collect the same information as in Wikipedia along with the Q code (identifier) of the individual.

Merging: we use this feature to improve the reliability of each information extracted from both sources. We develop and use a series of algorithms to i) come up with a relevant sample of humans, ii) eliminate duplicate biographies, iii) detect systematic errors contained either in Wikipedia or Wikidata and correct them.

Domains of influence

Discovery/Science
- Academia (research, historian, scientist, academic, …)
- Explorer (engineer, explorer, inventor, sailor, pioneer, …)

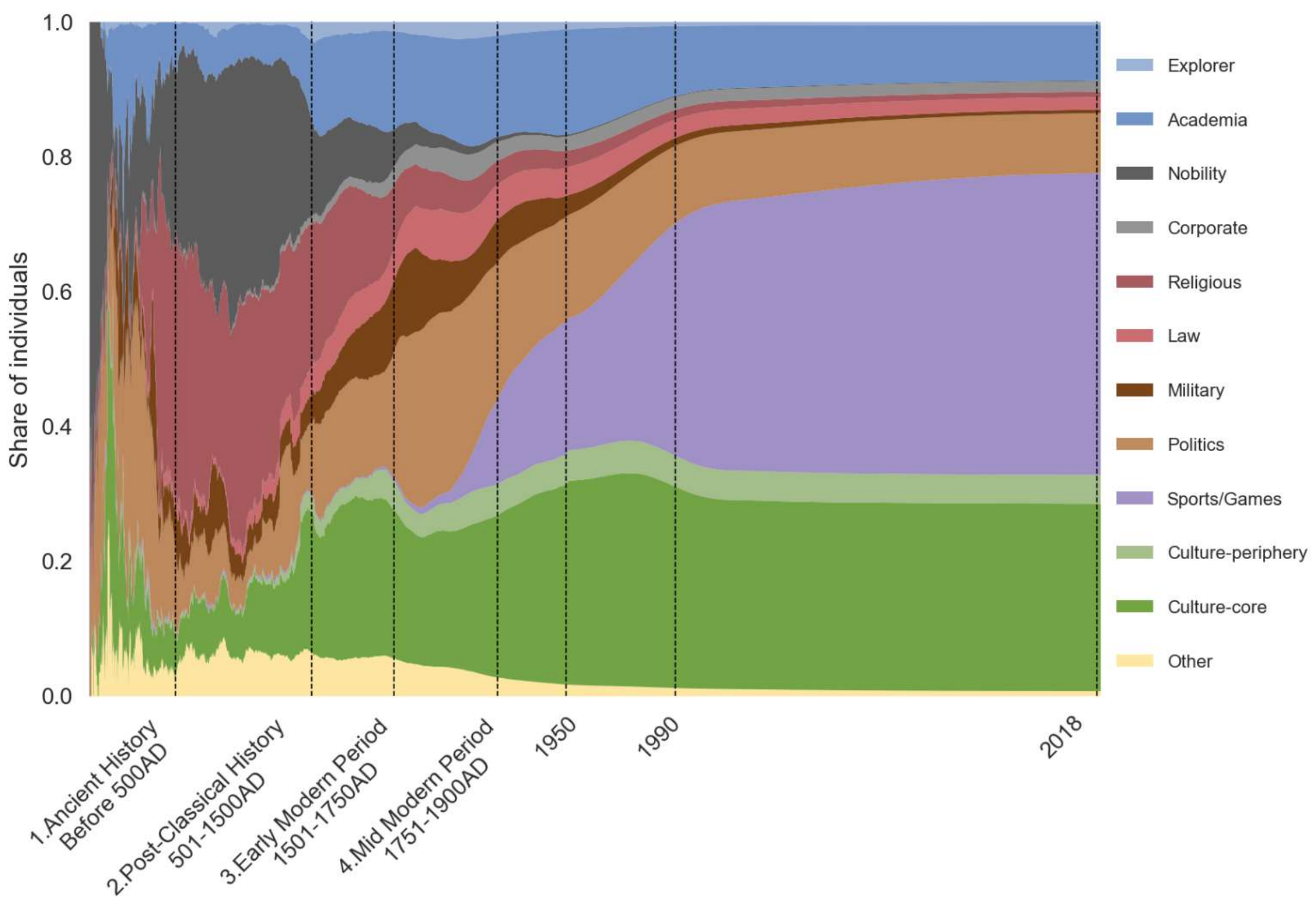
Culture
- Core (actor, writer, painter, singer, music, …)
- Periphery (journalist, architect, designer, presenter, ...)

Leadership
- Politics (politician, activist, trade unionist, etc.)
- Military (military, officer, commander, soldier, army, etc.)
- Law (lawyer, diplomat, judge, jurist, civil service)
- Nobility (aristocrat, noble, king, sovereign, monarch, etc.)
- Religious (priest, prelate, rabbi, missionary, bishop, etc.)
- Corporate leadership (business, entrepreneur, bank, etc.)

Sports/Games

Other: Worker (farmer, librarian, musher, ...) + Family (son, wife, father,...) + Misc. (esperantist, criminal, philanthropist,...)

Share of individuals, breakdown by domain of influence



Language groups

English language : individuals with at least one biography in the English edition of Wikipedia;

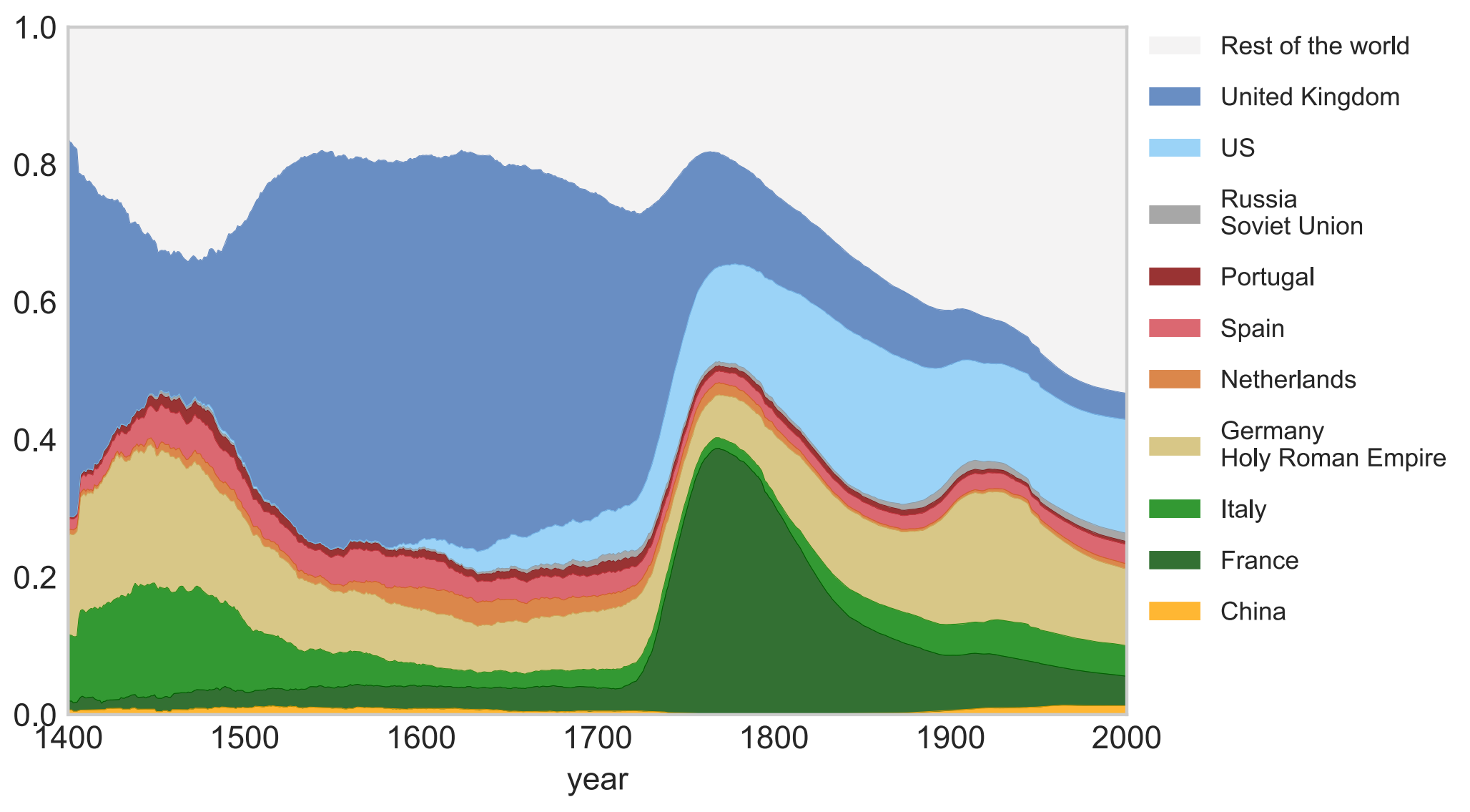
Western non-English : individuals with a Wikipedia biography in at least one of the Western languages but absent from the English edition.

Western, non-English editions are dominated by Culture and Politics and individuals from continental Europe (mostly Germany, France and Sweden) while the English edition is dominated by Sports, Culture and UK and US citizens.

Documenting key historical periods

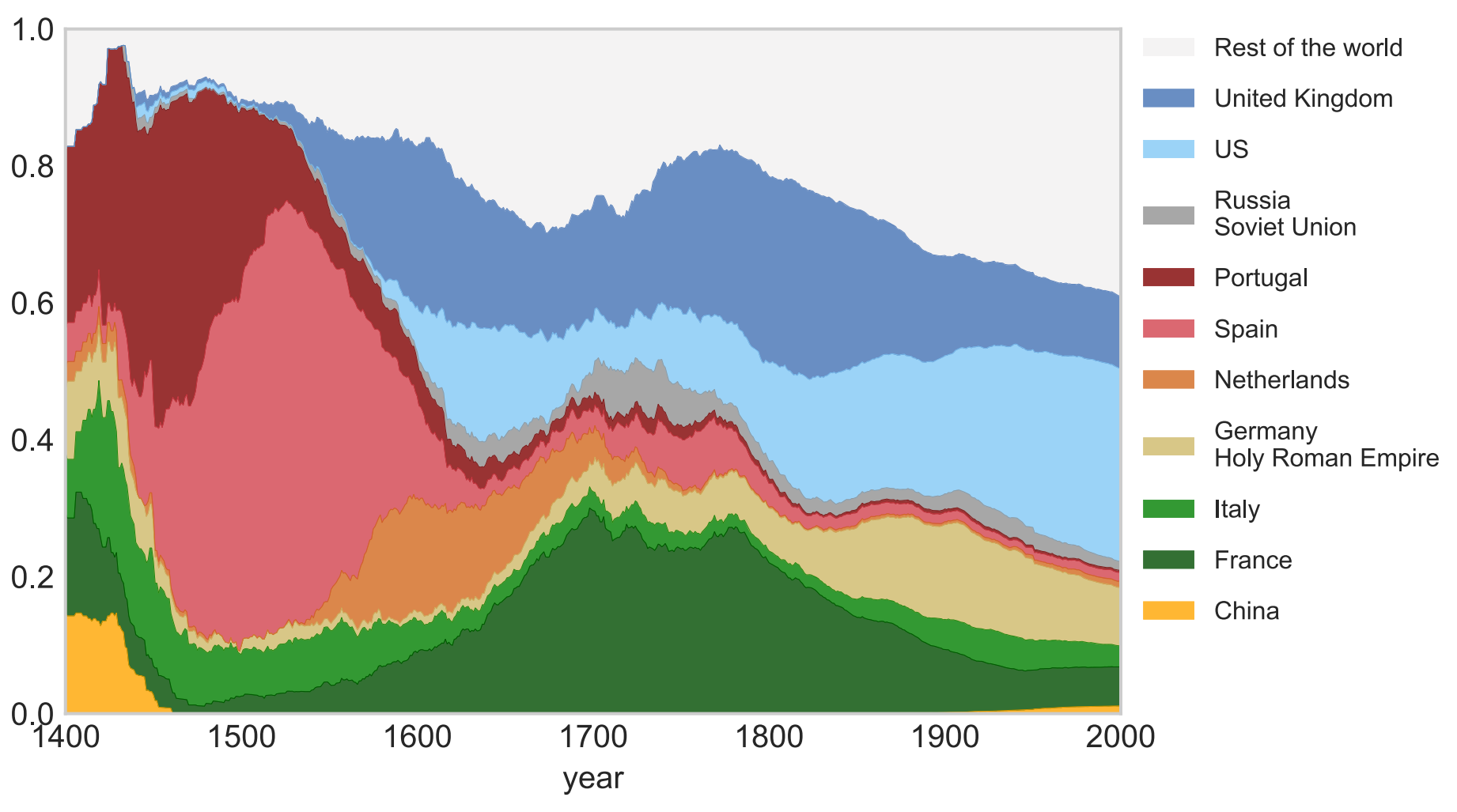
French and American Revolution (category: "politics")

The share of French and US politicians has been rising in a significant way around and after the end of the 18th century (creation of the new independent nation of the USA, French revolution)



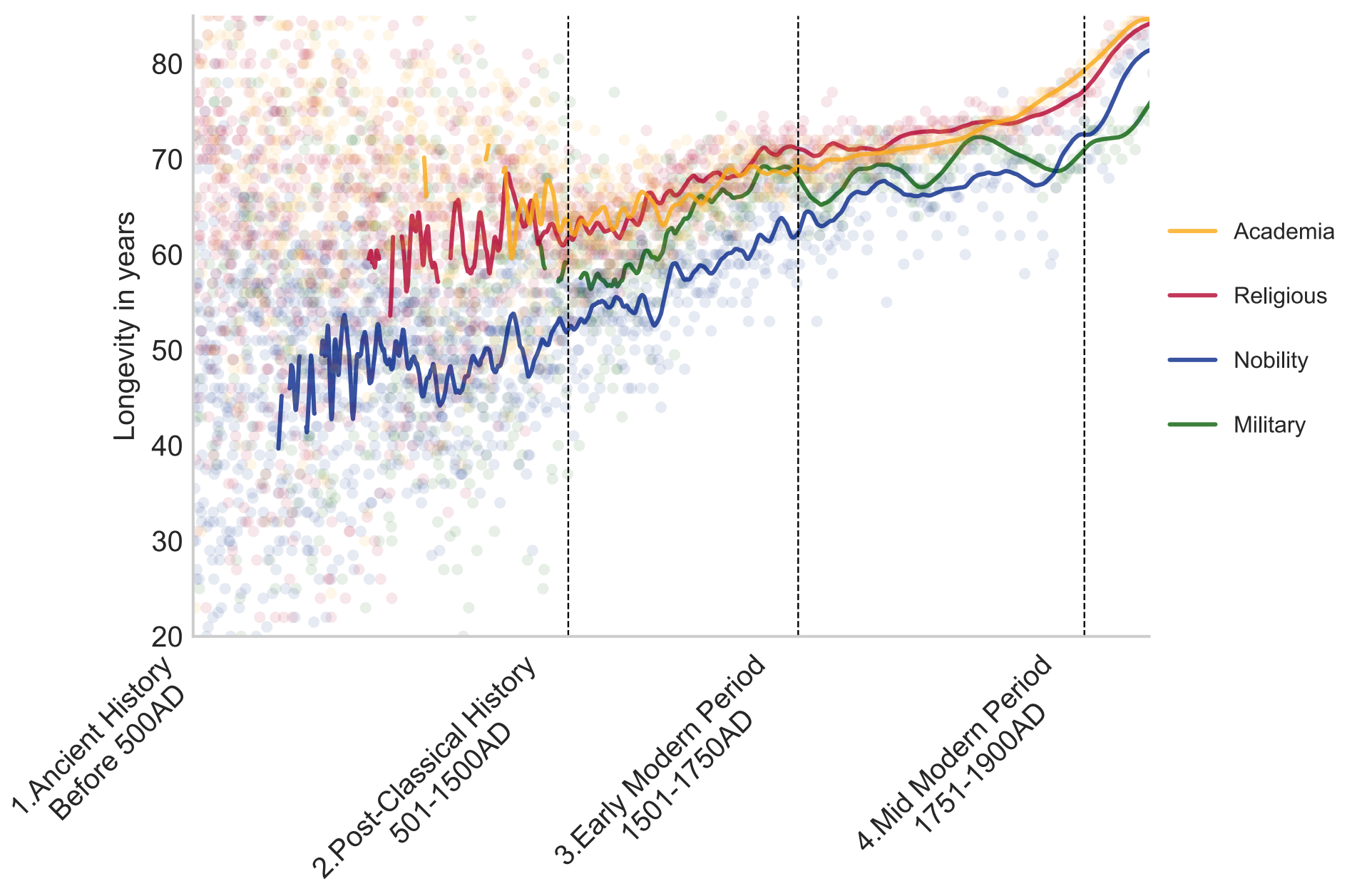
Age of discovery & European exploration (category: "explorer, Inventor, developer")

It illustrates the end of the Chinese exploration period which lasted until the 15th century followed by the European age of discovery and explorations conducted by Portugal and Spain in the 15th and 16th centuries.

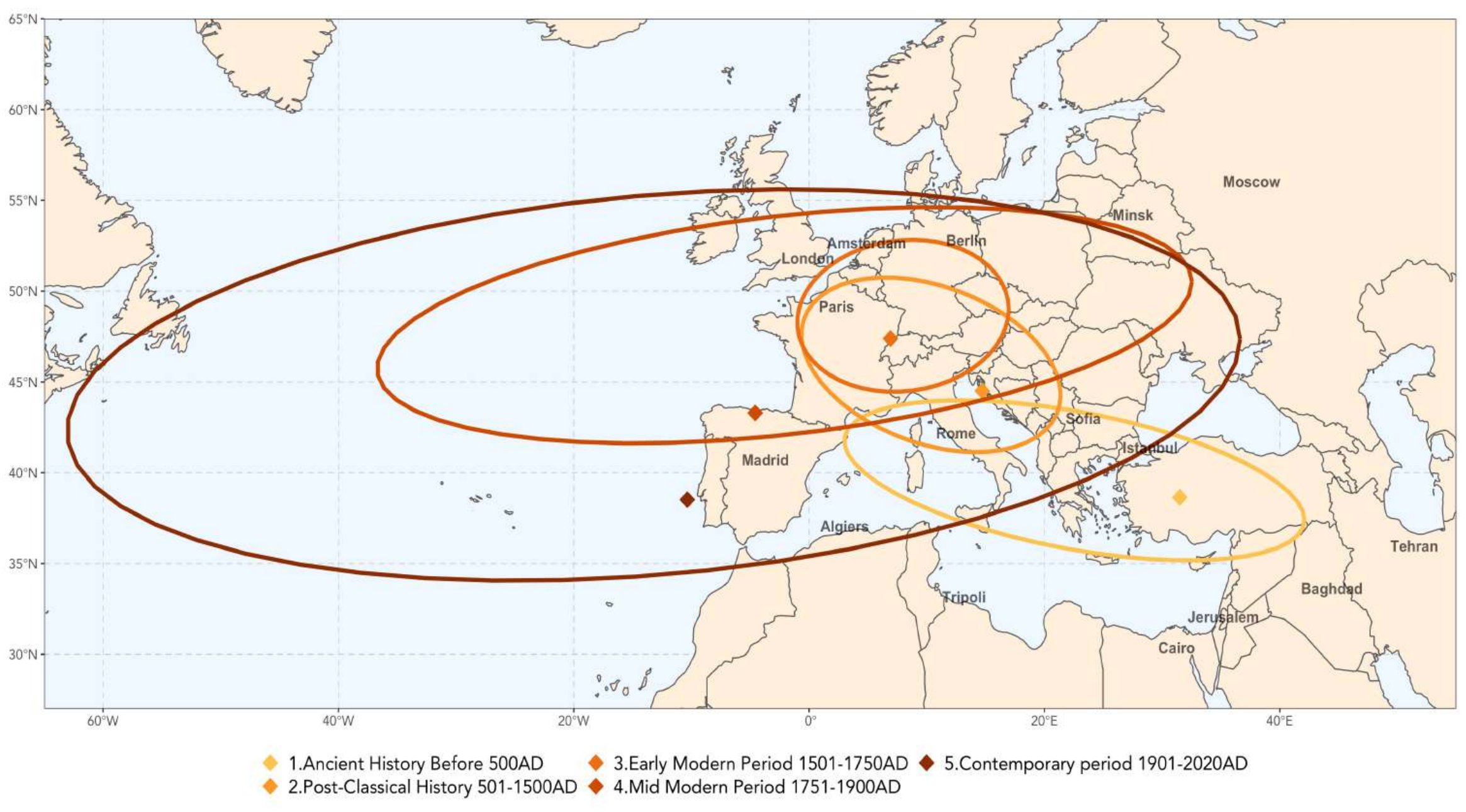


Median longevity…

… is lower for individuals in military and nobility, compared to academia and religion. Concerning nobility, the death of noble children drives down the median life expectancy of this category.



Covariance matrix ellipses and barycenters of birth places (Before 500AD to 1751-1900)


65°N
60°N
55°N
50°N
45°N
40°N
35°N
30°N
60°W
40°W
20°W
0°
20°E
40°E
Moscow
Minsk
Amsterdam
Berlin
London
Paris
Sofia
Rome
Istanbul
Madrid
Algiers
Tehran
Baghdad
Tripoli
Jerusalem
Cairo
1.Ancient History Before 500AD
2.Post-Classical History 501-1500AD
3.Early Modern Period 1501-1750AD
4.Mid Modern Period 1751-1900AD
5.Contemporary period 1901-2020AD

SciencesPo
LABORATOIRE INTERDISCIPLINAIRE
D'ÉVALUATION DES POLITIQUES PUBLIQUES

www.sciencespo.fr/liepp