A Brief History of Human Time. Exploring a database of " notable people "

Olivier Gergaud, Morgane Laouenan, Etienne Wasmer

SciencesPo
LABORATOIRE INTERDISCIPLINAIRE
D'ÉVALUATION DES POLITIQUES PUBLIQUES

LIEPP Working Paper

February 2016, nº46

A Brief History of *Human* Time

Exploring a database of "notable people"

Olivier Gergaud
KEDGE Business School and LIEPP, Sciences Po
olivier.gergaud@kedgebs.com

Morgane Laouenan
CNRS, Centre d'économie de la Sorbonne, U. Paris 1 and LIEPP, Sciences Po
morgane.laouenan@univparis1.fr

Etienne Wasmer
Sciences Po, LIEPP and Department of Economics,
etienne.wasmer@sciencespo.fr

www.sciencespo.fr/liepp

A Brief History of *Human* Time

Exploring a database of "notable people"

(3000BCE-2015AD) Version 1.0.1*

Olivier Gergaud†, Morgane Laouenan‡, Etienne Wasmer§

February 8, 2016

Abstract

This paper describes a database of 1,243,776 `notable people` and 7,184,575 locations (`Geolinks`) associated with them throughout human history (3000BCE-2015AD). We first describe in details the various approaches and procedures adopted to extract the relevant information from their `Wikipedia` biographies and then analyze the database. Ten main facts emerge.

1. There has been an exponential growth over time of the database, with more than 60% `of notable people` still living in 2015, with the exception of a relative decline of the cohort born in the XVIIth century and a local minimum between 1645 and 1655.

2. The average lifespan has increased by 20 years, from 60 to 80 years, between the cohort born in 1400AD and the one born in 1900AD.

3. The share of women in the database follows a U-shape pattern, with a minimum in the XVIIth century and a maximum at 25% for the most recent cohorts.

4. The fraction of `notable people` in governance occupations has decreased while the fraction in occupations such as arts, literature/media and sports has increased over the centuries; sports caught up to arts and literature for cohorts born in 1870 but remained at the same level until the 1950s cohorts; and eventually sports came to dominate the database after 1950.

*This text is updated on a regular basis, along with the changes and improvement in the database. Please check our website: http://www.brief-history.eu/ for the most recent version. Version 1.0 deposited on Jan. 31st, 2016. We thank Sarah Asset, Nicolas Britton, Jean-Benoît Eyméoud, Jessica Flakne, Simon Fredon, Valentine Watrin for outstanding research assistance, Atelier de cartographie and Medialab in Sciences Po for advice and discussions, and in particular Thomas Ansart, Benjamin Ooghe, Paul Girard and Patrice Mitrano. Having participated to other parts of the project, Meradj Aghdam, Mathis Forman, Florentin Cognie, Blaise Leclair, Charles Réveilleire and Lucy Rebel have our gratitude. We thank LIEPP's team, in particular Christelle Hoteit and Alexandre Biotteau, and Anne Le Page and the IT Department in Sciences Po. We benefited from very useful discussions with Pierre-Henri Bono and with the participants to the Seventh Conference on Cultural and Media Economics organized by the French Ministry of Culture and Communication, Sciences Po and KEDGE Business School, where this work was presented in September 2015. Financial support from the LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02) is gratefully acknowledged. In the paper, BCE refers to *Before Common Era* and AD to *Anno Domini*.

†KEDGE Business School and LIEPP, Sciences Po. Email: olivier.gergaud@kedgebs.com

‡CNRS, Centre d'Economie de la Sorbonne and U. Paris 1 and LIEPP, Sciences Po. Email: morgane.laouenan@univ-paris1.fr

§Corresponding author: LIEPP and Department of Economics, Sciences Po. Email: etienne.wasmer@sciencespo.fr

5. The top 10 visible people born before 1890 are all non-American and have 10 different nationalities. Six out of the top 10 born after 1890 are instead U.S. born citizens. Since 1800, the share of people from Europe and the U.S. in the database declines, the number of people from Asia and the Southern Hemisphere grows to reach 20% of the database in 2000. Coïncidentally, in 1637, the exact barycenter of the base was in the small village of Colombey-les-Deux-Eglises (Champagne Region in France), where Charles de Gaulle lived and passed away. Since the 1970s, the barycenter oscillates between Morocco, Algeria and Tunisia.

6. The average distance between places of birth and death follows a U-shape pattern: the median distance was 316km before 500AD, 100km between 500 and 1500AD, and has risen continuously since then. The greatest mobility occurs between the age of 15 and 25.

7. Individuals with the highest levels of visibility tend to be more distant from their birth place, with a median distance of 785km for the top percentile as compared to 389km for the top decile and 176km overall.

8. In all occupations, there has been a rise in international mobility since 1960. The fraction of locations in a country different from the place of birth went from 15% in 1955 to 35% after 2000.

9. There is no positive association between the size of cities and the visibility of people measured at the end of their life. If anything, the correlation is negative.

10. Last and not least, we find a positive correlation between the contemporaneous number of entrepreneurs and the urban growth of the city in which they are located the following decades; more strikingly, the same is also true with the contemporaneous number or share of artists, positively affecting next decades city growth; instead, we find a zero or negative correlation between the contemporaneous share of "militaries, politicians and religious people" and urban growth in the following decades.

There is currently a growing number of datasets allowing for the documentation of historical facts. A recent approach has focused particularly on historical individuals, who we call in this text `notable people`. This approach was pioneered by Schich et al. (2014). The authors automatically collected the years and locations of birth and death for 150,000 `notable people` in history using `Freebase`, a Google-owned knowledge database. de la Croix and Licandro (2015) built a sample of 300,000 famous people born between Hammurabi's epoch and 1879, Einstein's birth year from Index Bio-bibliographicus Notorum Hominum, to estimate the timing of improvements in longevity and its role in economic growth. Recently, Yu et al., (2016) also used `Freebase` and assembled a manually verified dataset of 11,341 biographies existing in more than 25 languages in `Wikipedia`. Our paper extends these approaches. We compile the largest possible database of `notable people` rather than focusing only on "very famous" individuals, because we are ultimately interested in detecting the statistically significant local economic impact of these individuals. It actually turns out that weighting individuals with measures of their impact does not make a big difference, which *ex post* justifies our collection of information on hundreds of thousands of lesser known artists, business people and local rulers, famous enough to have been listed and described somewhere on the internet or in various rankings, but yet left out of the vast majority of internet sources.

To this end, we use two different, yet complementary approaches to obtain names of and information on `notable people`. One is also based on `Freebase`, and allows us to collect information at a large scale for 938,000 individual profiles over 4,000 years of human history, after a careful examination of homonyms and the elimination of duplicates. We refer to this method as "top-down", since the information on names is centralized in `Freebase`. The second method is based on a systematic search from various categories in `Wikipedia` pages to identify `notable people` and is refered to as a "bottom-up" approach. This

results in a list of about 1 million individuals that considerably overlaps with the "top-down" approach, and eventually adds another 280,000 names to the `Freebase` results.

These individuals are then matched with their respective `Wikipedia` biographies in English, from which we extract a large amount of biographical information through a careful and manually verified semantic analysis. We categorize people according to gender, nationality, and their three main activities/occupations according to an *ad hoc* classification system consistent throughout history. Other key information such as birth year, death year, birth place, place of death and most importantly all geographical linkages (`GeoLinks` thereafter) mentioned in the biography in between birth and death, have been systematically collected where available. A total of 3.5 million geographical linkages have been gathered and analyzed.

We also attempt to measure these individuals' impact on various economic outcomes. The visibility of each individual's page is probably a better proxy of his or her impact than the number of pages viewed over the past years (the alternative approach pionneered by Yu et al., 2016). Therefore, we use a simple impact measure compiled using a combination of the number of words in the `Wikipedia` page and the number of languages into which the page has been translated. The implied ranking is disclosed in the text and its Appendix, both in an overall ranking and in rankings by categories. Alternatives are explored and correlations between them analyzed.

We then match `notable people` with cities using a unique global historical population database composed of new Census data from 17 countries on 5 continents, the Urbanisation Hub-Bairoch-Bosker city population data between 800AD - 1800AD, the Lincoln population data 1794 - 2005 for 24 other cities in the world, and the United Nations (UN) population database since 1950. For each country and within each period, we select the 30 largest cities at a time for countries with a landsize of less than approximately 300,000 square kilometers and for the 50 largest cities for larger countries with a landsize larger than 300,000 square kilometers. The landsize threshold is more time invariant than any other population criterion and the number of cities retained permits that 50% of 3.5 million locations assembled in the database over the period 800AD-2015AD lie within 50km of the nearest city. In the last period of the sample (1950AD-2015AD), the UN population database even allows us to have 50% of the locations within 13km of the nearest city.

In this paper, we carefully document the data search and procedures to extract the relevant information. We then explore the datasets, provide a number of descriptive stylized facts and finally provide descriptive, non-causal associations of the role of `notable people` on city growth.

1 Data collection

1.1 A list of individuals from `Freebase` then matched with `Wikipedia` pages (*"top-down" approach*)

We exploit information on `Freebase` as documented in Schich et al. (2014). `Freebase` is a Google-owned knowledge database providing a list of notable individuals known to have existed. This list comes from a variety of web sources, including `Wikipedia`, the `Internet Movie Database`, `Allocine`, and others. The link we refer to is: https://www.freebase.com/people/person?instances=. The procedure includes several steps.

1. From `Freebase`, we collect both the full name (`first [middle] last names`) of each individual, and

his/her `birth date` (if available). As of January 2015, when we accessed the list, `Freebase` contained precisely 3,440,707 `notable people`. However, many names were duplicates, or mistakenly added. A subset of nearly 407,000 irrelevant observations, corresponding either to pure html codes, movie titles, ships such as HMS Titanic, names in non-Latin alphabets (overall, this eliminates around 3,000 names that appear in Chinese/Korean characters, in Greek or Cyrillic alphabets) or even numbers instead of full regular names have been identified and then automatically dropped out of the initial dataset. This step leads to a sample of 3,033,469 `notable people`.

2. We then match these 3 million names with their corresponding `Wikipedia` pages in English. The matching process was more easily facilitated when using the birth date, where available, especially for homonyms. Overall, 46% (1,391,718 observations) include a birth year. The matching process achieved a lower success rate for people with no such birth information, i.e. when the matching process was based on the full name only (see next point). Missing birth dates represent 54% of the sample (1,641,751 observations).

3. There are several homonyms in the database. When `Wikipedia` detects homonyms it either provides a list of names with birth dates[1] or directly takes the user to the page of the most famous homonym e.g. for Ray Charles (https://en.wikipedia.org/wiki/Ray_Charles), via the following sentence and link : "*For other uses, see Ray Charles (disambiguation).* In this second case the link refers to a list of homonyms with a link to each one's respective page. In both cases, in order to capture as many homonyms as possible from `Freebase` we match both the full name and the birth year. For homonyms with no birth information in `Freebase`, we only consider the page corresponding to the most famous homonym instead of collecting all pages mentioned in the list described before (first case).

Among the individuals with birth year information, approximately 30% do not have a `Wikipedia` page in English, 868,266 have a unique link to a page in English, while 154,349 are not matched. Among these pages, 76,000 must be disambiguated as they contain more than one link. This leads to a first group of names. Among individuals with no birth year information, 313,308 come up with a unique link to a `Wikipedia` page in English, but only 106,646 refer to an individual (and not to a list of homonyms, concepts, etc.). Of these, 37,854 are not duplicates with respect to the first group detailed above (with birth year information) and are therefore included in the final sample.

At the end of the process, 964,245 individuals are matched with their `Wikipedia` biography, which we downloaded as of December 2015.

1.2 A list of individuals using `Wikipedia` categories (*"bottom-up" approach*)

To be exhaustive and consistent, we also use `Wikipedia` directly, which classifies most individual pages by birth date. A secondary independent dataquest is run automatically from birth dates between 1500BCE to 2015AD; a complete list of categories is assembled; virtually all of these categories include additional lists of people. Each of these is scanned to obtain more names with a page in English. We obtain a total of 1,176,812 names including 897,310 names in common with names identified using the "top-down" approach and 279,531 distinctly new names.

We then merge the two databases ("top down" and "bottom up") using the url of each individual (`https://en.wikipedia.org/wiki/[first name+last name]`. Figure 1 describes the process and the final outcome. The final database includes 1,243,776 `notable people`.

[1] E.g. for John Martin we get : https://en.wikipedia.org/wiki/John_Martin

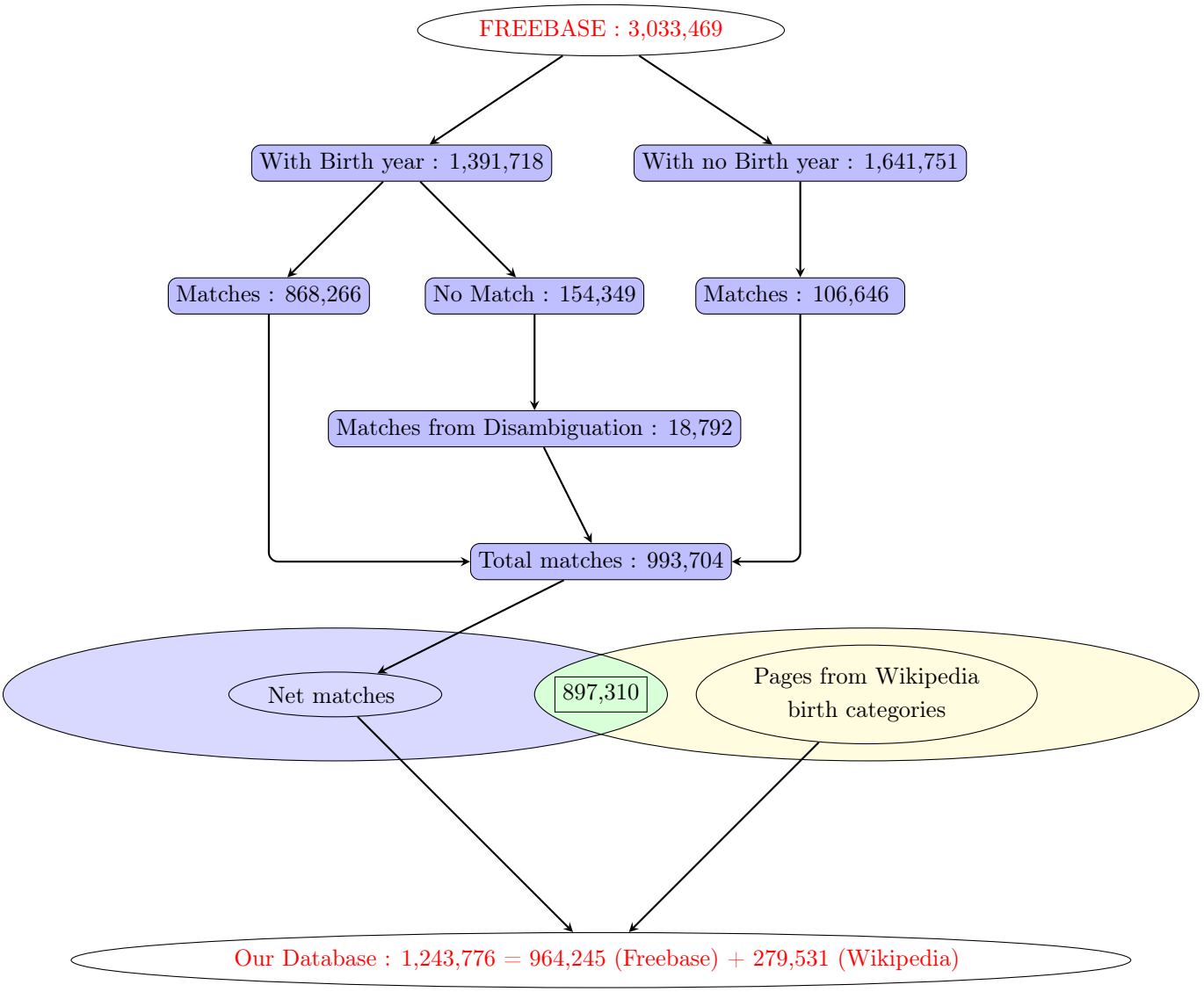


Figure 1: Organization chart

1.3 Individual characteristics

In this subsection we analyze the `source code` of each `Wikipedia page` to extract basic information about dates and locations for birth and death, occupations, citizenship and gender. These individual characteristics are either identified from the `Infobox` (fixed-format table at the top right-hand corner of biographies), from the `Abstract`, or from the `Categories` section. 398,830 (36.05%) individuals have no `Abstract` and therefore we extract information from the `Main text`.
Figure 2 shows the different parts of the `Wikipedia` page of Ray Charles. See his full page in the Appendix.

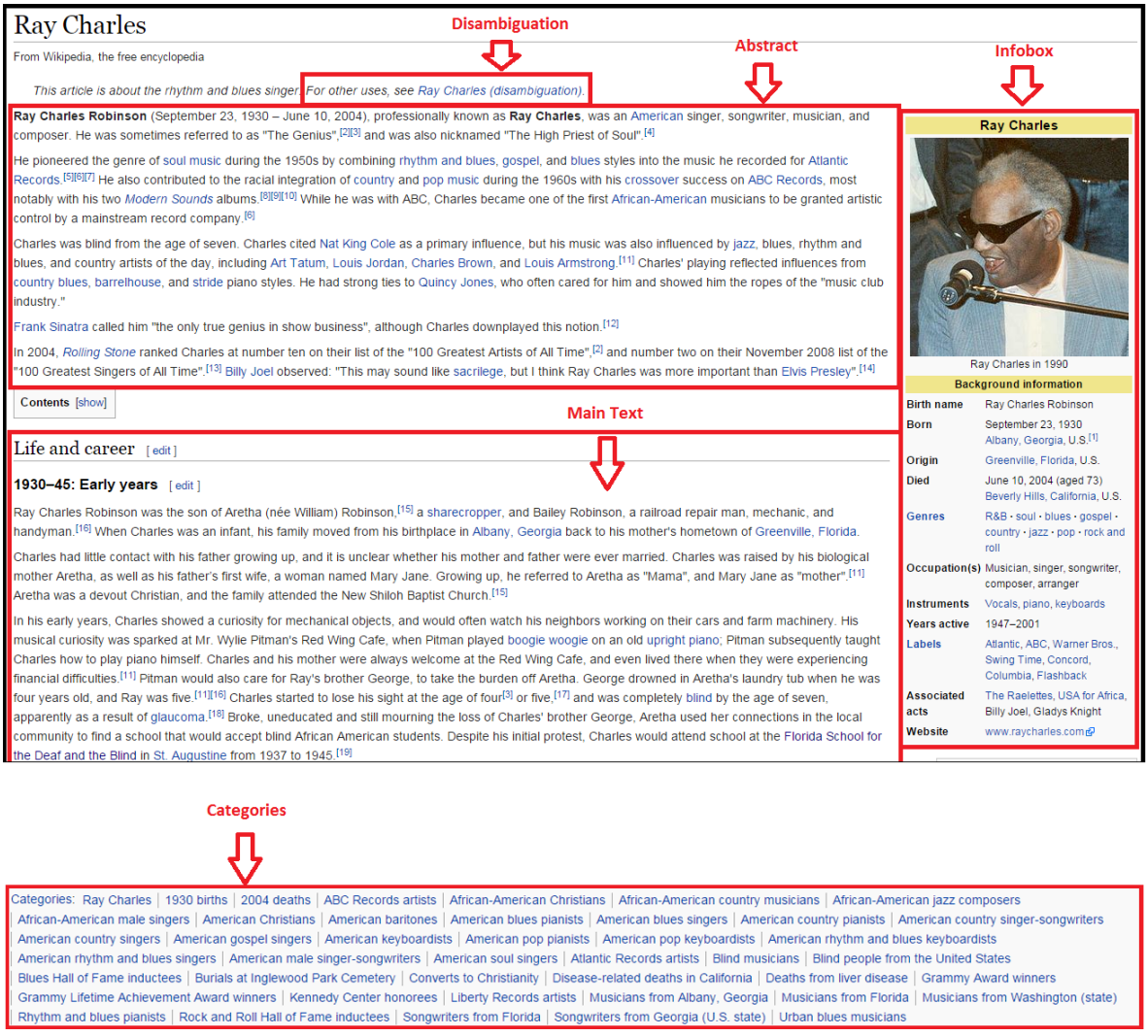


Figure 2: Example of a `Wikipedia page`: Ray Charles

Birth and Death Dates Information on an individual's date of birth and death is typically available in three different sections of the `Wikipedia` biography: i) `Infobox`, ii) `Abstract/Main text` and iii) `Categories section,` which is located at the bottom of the page.
We therefore get a maximum of three birth dates per individual of which 1,063,637 come from the `Categories section`, 669,405 were found in the Infobox and 951,147 were extracted from the `Abstract/Main text`. Some are misreported but the alternative birth dates allow for a correction: among observations without missing dates, the concordance rate between birth dates coming from the same biography is at around 97.3%. Similarly, we found 498,065 dates of death in the `Categories section`, 232,584 in the `Infobox` and 383,088 in the `Abstract/Main text`. The concordance rate between dates of death coming from the same biography is at around 98.4%.
We end up with 1,073,585 birth years and 499,980 death years. Missing death years correspond either to an unreported death year or to an individual still living in 2015 (a large fraction of our database : 59%). Note also that we have 144,804 individuals with no birth information available regardless of the three sources used. Among those individuals, 21,581 have information available on their death.

Places of Birth and Death Places of birth and death may both be found in the `Infobox` and/or in the main body of the text (`Main text`). We first analyze the information coming from the `Infobox` and then consider the `Main text` as an alternative source of information when either the `Infobox` is missing or the relevant information has not been detected in the `Infobox`. In this case, we use keywords such as "born in/at" or "died in/at" to find place of birth and death. Most locations of an individual's birth and death are associated to a latitude and a longitude (hereafter we use the word geocoded), which we then extract.

Citizenship/Gender Information about citizenship is usually present in the `Infobox` and in the `Abstract/Main text`. We use both sources in order to minimize the number of missing values. On rare occasions, when there is more than one citizenship/country mentioned in the `Abstract`, we consider the citizenship appearing first as that individual's citizenship. After thorough manual verification, it appears that this information always appears first in the `Abstract section` (as illustrated below in Ray Charles' biography).
For gender we also use the `Abstract/Main text` and check for the presence of pronouns (he/she) and possessive adjectives (his/her) in the text. We consider a person to be female (male) if "she"/ "her" ("he"/ "his") is found in the summary. In case we detect both masculine and feminine pronouns or possessive adjectives, we select the first pronoun that appears as the one identifying the person's gender. This method did not prove efficient for short biographies which do not contain any pronouns or possessive adjectives. Overall, women account for only 15.7% of the sample, while men represent the largest share (78.5%) and we failed to identify the gender information for only 5.8% of all biographies analyzed. However, a careful visual inspection of this latter category indicates that missing genders are predominantly males.

Occupations We determine occupations by locating linking verbs such as "was a"/"is a"/"was the"/"is the" in the `Main text`: for instance, "*Ray Charles was an* ***American*** *singer, songwriter, musician, and composer*". Occupations are then grouped into 6 categories: Academics (studies, education), Entertainment (arts, literature/media, sports), Entrepreneur (business, inventor, worker), Family, Governance

Occupation A1	Frequency	Share
Academics (studies, education)	110,610	9.0
Entertainment (arts, literature/media, sports)	736,626	60.0
Entrepreneur (business, inventor, worker)	59,028	4.8
Family	11,813	1.0
Governance (law, military, religious, politics, nobility)	285,514	23.3
Other	23,712	1.9
Total	1,227,303	100.0

Table 1: Occupations (level of aggregation A - six categories)

Occupation B1	Frequency	Share
Arts	241,042	19.6
Business	48,974	4.0
Education	36,542	3.0
Family	11,813	1.0
Inventor	4,783	0.4
Law	28,585	2.3
Literature/media	103,421	8.4
Military	40,516	3.3
Nobility	20,013	1.6
Other	23,712	1.9
Politics	160,696	13.1
Religious	35,104	2.9
Sports	392,163	32.0
Studies	74,068	6.0
Worker	5,271	0.4
Total	1,226,703	100.0

Table 2: Occupations (level of aggregation B - 15 categories)

(law, military, religious, politics, nobility) and Other. Tables 1 and 2 provide a few summary statistics of the entire database. See the full list of occupations collected and sorted in the Appendix.
Table 3 summarizes the information that has been collected so far and informs us about the location where the information was found in the `Wikipedia` biography.

Visibility Factors Various "influence weights" are built, based on specific page characteristics combining the length of the page (in words), the number of translations of the page and some additional information such as the number of footnotes, categories, headlines, links, etc. Futher details about these are provided in Section 2 below. As argued in the introduction, many other sources, depending on the type of study envisaged can serve as a proxy for the visibility factor. Our measures fit well our purpose, which is to detect the local economic impact of the individuals present in the database.
The number of pages viewed in recent years is a potential indicator that has been favored by Yu et al. (2016), and would be very useful if we were interested in studying contemporary individuals (living artists, athletes). The number of internet pages (backlinks) linking to a `Wikipedia` biography could also be an indicator. The latter information would be interesting *per se* as a measure of the relevance of these biographies. For our purpose, the information on the length and complexity of a `Wikipedia`

Individual variables	Location within the Wikipedia biography	Available
Birth & Death Dates	Abstract/Main text + Categories + Infobox	1,073,585 & 499,980
Birth & Death Places	Main text + Infobox	842,926 (67.77%) & 235,514 (46.25%)
Occupation	Abstract/Main text	1,183,971 (95.19%)
Gender	Abstract/Main text	1,149,899 (92.45%)
Citizenship	Abstract/Main text + Infobox	1,090,190 (87.65%)

Shares for death dates and death places are computed using the number of dead people only, i.e. 509,181 observations.

Table 3: Number of individuals with available information

biography and the number of links leading outwards to the Internet is still the most appropriate: we presume that individuals who have a Wikipedia biography and another biography elsewhere on the internet have had precisely the impact we attempt to detect. Of course, we may want to compare an individual's visibility relative to others in their respective birth cohort where relevant.

1.4 Geographical Wikipedia linkages (GeoLinks)

An important improvement of the present work compared to other articles quoted above is our in-depth and careful analysis of GeoLinks present in individual Wikipedia biographies. This analysis provides detailed and reliable information about these different places where famous people live (lived) and/or interact with (interacted with) over the course of their lifetime[2].

To collect such information we copy all hyperlinks found in each page and make a distinction between links found in the abstract and links coming from the Main text. Hyperlinks, by definition, lead to other Wikipedia Pages (Wikilinks), which we extract and parse. Wikilinks with geographical coordinates (longitude/latitude), are potential GeoLinks such as Albany, Georgia where Ray Charles was born.

We discard countries, regions or provinces that are locations with geographical coordinates as our goal is to match notable people with places at the city level (see below). There are also pages containing coordinates that do not correspond to locations, but to individuals. These pages provide the coordinates of their resting/burial place. For instance, the wikilink Ronald Reagan appears in Charles' Wikipedia page. Charles performed for Reagan's second inauguration in 1985. Ronald Reagan is considered by our code as a location because his page provides the coordinates of his resting place (Ronald Reagan Presidential Library, Simi Valley, California, 34.25899°N 118.82043°W). We can identify these wikilinks that contain coordinates but which are not locations and therefore exclude them.

A GeoLink may not necessarily point to places where famous people at some point moved to, lived in or even just visited. After careful verification of hundreds of cases[3], it appears that a significant fraction of these GeoLinks refer to parents/family, or to a place where they lived, originated from, etc. This information might be relevant for some uses, but is not relevant in analyses of an individual's direct impact, so we keep both but separate them out, distinguishing between two types of GeoLinks: those having a connection with family and those with no obvious link to family, using keywords such as

[2]We will use thereafter either the present or preterit tense to talk about present and past linkages respectively.

[3]These verifications are available from authors upon request.

Wikilinks	GeoLinks	Countries/Regions	Resting Place	Family	Post-mortem
7,184,575	4,878,455	1,829,400	32,915	398,3331	363,318
	(67.90%)	(25.96%)	(0.47%)	(5.65%)	(5.06%)

Table 4: Number of identified GeoLinks

"sister", "father", "brother", etc. contained in the same sentence to allocate them into the first category. Other GeoLinks refer to post-mortem events and can safely be dropped out of the sample. These are quite frequent for artists (e.g. links to museums where exhibitions of their work take place *post-mortem*) or scientists (e.g. honorary awards, buildings named in his/her honor). We consider a location to be *post-mortem* whenever a GeoLink is pointed out after an individual's date of death and is present in a sentence where keywords such as "died/buried/..." are present.

Table 4 shows that among all 7,184,455, Wikilinks extracted 26% correspond to either countries or regions, 0.5% refers to a person instead of a location, 5.7% are places visited by family members and 5.1% are *post-mortem* locations. These cases are not mutually exclusive. In all, we detected 4,878,455 proper and usable GeoLinks.

Places are either located in the Abstract or in the Main text. Therefore, in order to keep a chronological path of places visited, we keep either the Abstract or the Main text depending on the number of places contained in both sections. We extract places from the section containing the largest number of GeoLinks. The Abstract is chosen arbitrarily in the case of a tie. On average, individuals have 5.7 GeoLinks in the Main text and only 2.4 in the Abstract.

We detail here the places visited by Ray Charles (1930-2004) as it was identified by our automatic detection procedure. Charles was born in Georgia and grew up in Florida before reaching cities like Seattle, Philadelphia, New York and Los Angeles. Along with Table 5 providing the GeoLinks, Figure 5 in Appendix shows the screenshots of the different sections of his Wikipedia page including GeoLinks. Figure 3 shows the map of places visited by Ray Charles according to his Wikipedia page. It is important to report that not all GeoLinks refer to city names. Some point to cultural events, museums, opera houses, theaters, universities, schools, stadiums or sports events such as Olympic games, etc.. Those links have coordinates and are added to GeoLinks. In this example, Ray Charles had two contacts with the city of St. Augustine, Florida, through a school first (Florida School for the Deaf and the Blind) and a radio station (WFOY) based in the same city. This is important information to consider as sampled individuals have had contact with these cities through these institutions.

Order of appearance in the biography	`GeoLinks`
1	Albany, Georgia
2	Greenville, Florida
3	St. Augustine, Florida
4	WFOY (radio station), St. Augustine, Florida
5	Jacksonville, Florida
6	Ritz Theatre (Jacksonville)
7	LaVilla, Jacksonville, Florida
8	Orlando, Florida
9	Tampa
10	Seattle, Washington
11	Overtown (Miami)
12	St. Petersburg, Florida
13	The Apollo Theater
14	Uptown Theater (Philadelphia)
15	The Newport Jazz Festival
16	Beverly Hills, California

Table 5: `GeoLinks` of Ray Charles

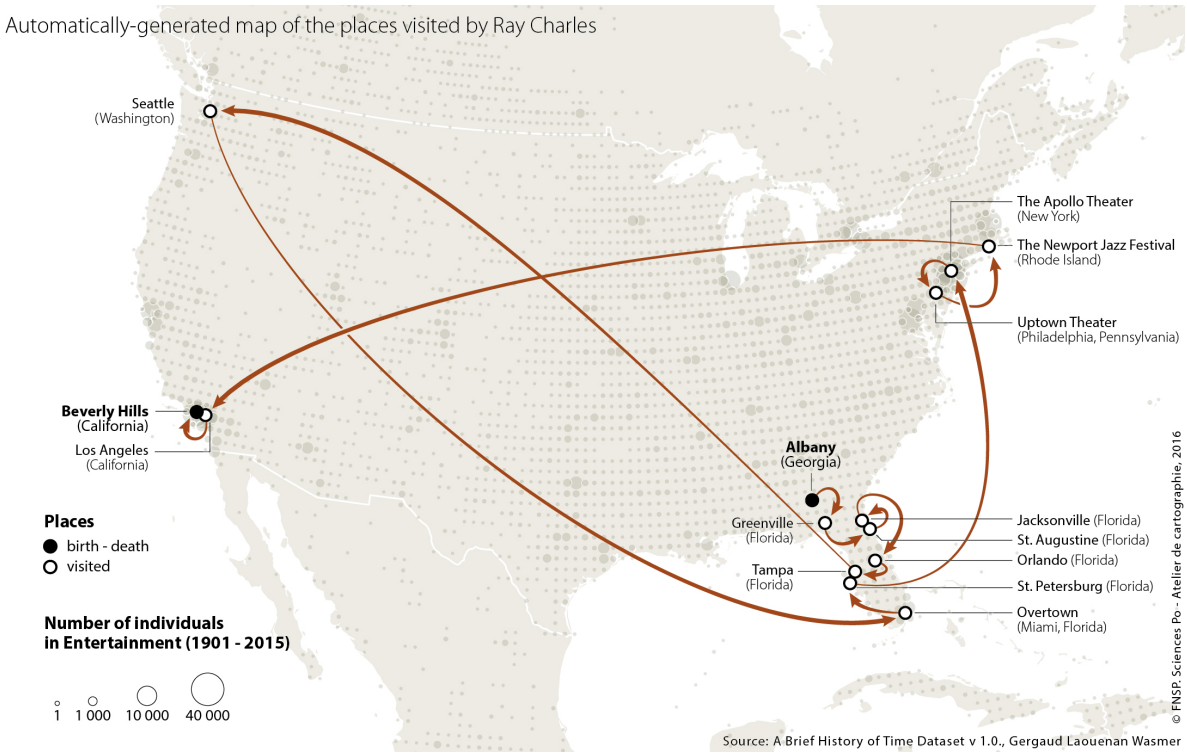


Figure 3: Automatically identified `GeoLinks` of Ray Charles (1930-2004)

We show here, as another example, the trajectories of a person known for having been very mobile over the course of his life: Desiderius Erasmus (1466-1536AD) in Figure 4. Erasmus was born near Rotterdam (Netherlands: 1) with some controversy about whether he originated from Gouda (Netherlands: 2). Erasmus has an indirect (weaker) link with at least two different `GeoLinks`. First, with the city of Zevenberger (Netherlands: 3) which is the place of birth of his grandfather and second with the city of Deventer (Netherlands: 4) where his oldest brother went to a famous school. Both places are family-related `GeoLinks` and are therefore in blue on the map. As previously explained, these two locations have been dropped from the list of `GeoLinks`. According to `Wikipedia`, the connection between Erasmus and Nuremberg (Germany: 5) is through a Portrait of him by Albrecht Dürer in 1526 which he engraved in that city. Next, Erasmus was tutoring in Paris (France: 6). Paris was mentioned no less than eight times in Erasmus' biography. Then he moved from Paris to Cambrai (France: 7) when he became a secretary to the Bishop of the city. Then he returned to Paris to study there at the University (France: 8). After Paris he moved to Leuven (Belgium: 9) where he became a lecturer at the Catholic University. His next move was from Leuven to Cambridge (United Kingdom: 10) where he held the position of Professor of Divinity at the University from 1510 to 1515. In 1506, he graduated as Doctor of Divinity from the Turin University (Italy: 11) and also worked part time as a proofreader at a publishing house in Venice (Italy: 12). Erasmus then emigrated from Venice to Basel (Switzerland: 13), a city that he left in 1529 to settle in Freiburg im Breisgau (Germany: 14) where he eventually died at the age of 69. Finally, we report on Figure 5 the birth to death trajectories on a smaller sample (1/20th of individuals with initials A and some B) and the barycenter of individuals in the database by time period and their dispersion, from individuals born before 500AD (yellow ellipses) to the most recent period (darker ellipses). Ellipses are constructed from the standard deviations of longitude and latitude. One can observe first, on top, the concentration of locations around an axis Europe-North America and second, in the bottom part of the Figure, the slow movement of the sample from the Middle-East to Western Europe with a barycenter in France during the XVIth and XVIIth centuries and towards North America, returning to the South-East in the last two centuries with the emergence of Africa and Asia (see Section 2). Coïncidentally, in 1637 (year of "Discours de la Raison by René Descartes), the exact barycenter of the base was actually in the small village of Colombey-les-Deux-Eglises (Champagne), where Charles de Gaulle bought his family house and died. Since the 1970s, the barycenter oscillates between Morocco, Algeria and Tunisia.

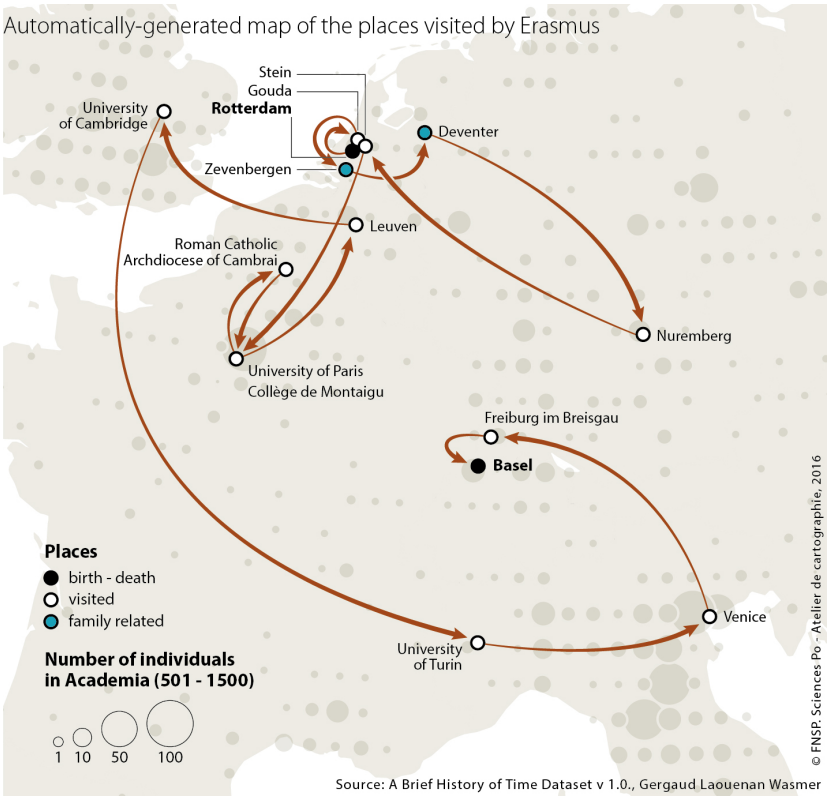


Figure 4: Automatically identified GeoLinks of Desiderius Erasmus (1466-1536AD)

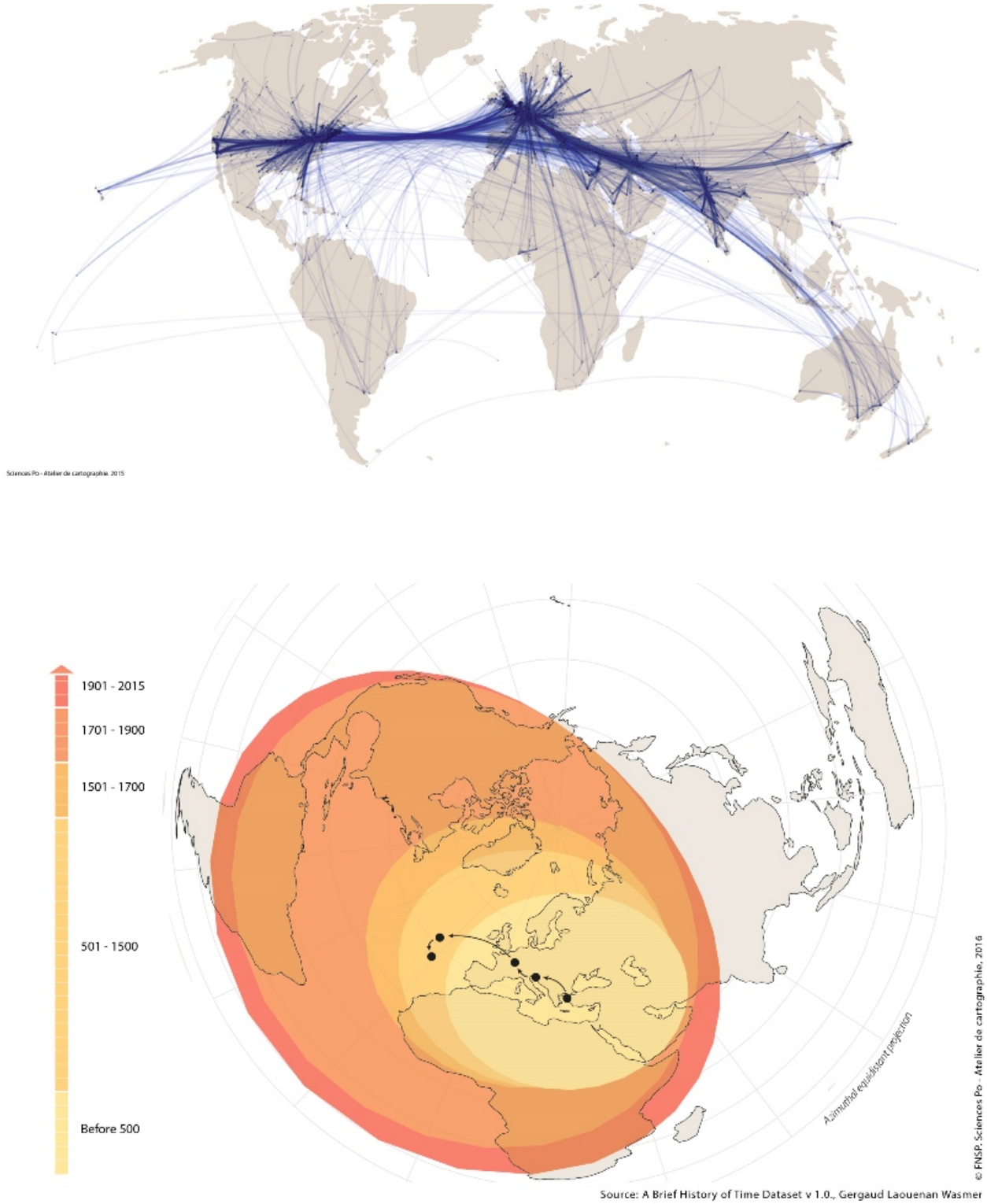


Figure 5: Top: birth to deaths trajectories of 1/20th of individuals. Bottom: barycenter of individuals in the database by time period and their dispersion from individuals born before 500AD to individuals born after 1900AD.

1.5 Matching procedure of GeoLinks to Time Periods

For subsequent descriptive analyses, we often group individuals into aggregate time categories. At the highest agregation level we use 5 large historical periods (<500AD; 501-1500AD; 1501-1700AD; 1701-1900AD; 1901-2015AD); see for instance the different maps of Europe and the world in Appendix A. We also use, in the econometric analysis of later sections, 66 Time Periods of varying length constructed as follows: individuals born before 1AD are divided into three groups, due to sample size (born before 500BCE, born between 501BCE and 250BCE, born between 249BCE and 0); periods of 50 years between 1AD and 1700AD; periods of 25 years between 1701AD and 1800AD and periods of ten years between 1801AD and 2015AD. These varying lengths of time capture the "accelerating history" phenomenon and reduce the disparities in the number of individuals across time periods.

The most important methodological choice here is to attribute to each GeoLink one or two of these Time Periods, as follows. Information on birth and death allows us to calculate the lifespan of each individual. We also estimate the mean lifespan according to each birth year or each Time Period, for both male and female individuals. When either the birth or death information is missing, we therefore impute lifespan based on the relevant Time Period of the gender category of the individual. Once done, we count the number of GeoLinks for that individual and use it to calculate the steps of a grid of the entire lifespan, as a proxy for the time elapsed between two successive GeoLinks. For people still alive in 2015, we use their current age in 2015 rather than their lifespan. Then, if a GeoLink is estimated to start in year a and end in year b, we assign that GeoLink to Time Period n if any year in the interval (a, b) intersects that Time Period .

The first GeoLink of a given individual should in principle be his or her birthplace and the last GeoLink should be his or her place of death. This might not always be the case, however, and in hundreds of manually verified cases, the GeoLinks were sometimes indicating high-schools or other places. For these people, we therefore add a first GeoLink with the birthplace when available in the Infobox, or a last GeoLink with the place of death when available from the same source.

1.6 Validity of the extraction procedure

The extraction procedure presented above has been checked by a series of manual verifications based on 4 randomly selected sub-samples of extracted biographies (853 in total). Both the individual characteristics (birth and death dates, places of birth and death, citizenship, gender, occupation) and GeoLinks automatically extracted by our code have been cross verified. At each round of verification, the code was modified and re-run on the entire sample. We present in Table 6 detailed statistics about the different iterations. The rate of errors on variables is generally decreasing (and ends up being lower than 7%) and the one on GeoLinks is limited (lower than 10%). Further improvements of the code are envisaged to improve this systematic detection procedure.

Data scientists and statisticians have introduced the concept of random dataset in which any information is subject to error and is potentially weighted by a probability of error. This will be our next step in future versions of the work. In the remaining sections we present some stylized facts as well as some correlations obtained between the various statistics generated by our Wikipedia extraction procedure and city population data as another *ex post* check of the quality of the database.

	GeoLinks	Individuals	% Residual errors on individual characteristics	% Residual errors on places
Round 1	452	114	24.5	20.1
Round 2	711	113	35.3	14.3
Round 3	849	152	13.8	11.4
Round 4	1316	81	6.17	9.12

Table 6: Manual verifications: performance statistics in each round

2 Facts on notable people

2.1 Size of cohorts in the sample

The sample is unbalanced across birth years, with an exponential growth of the sample size over time. The first individual in the database was born in 2285BCE, but the median individual was born in 1943 AD (column I). Only 1% (1099 individuals) were born before 1330. The composition of individuals not selected from Freebase ("top-down" approach) but from Wikipedia ("bottom-up" approach) is not very different in terms of years of birth. The size of a cohort for a given year in the database varies from 0 or 1 to as much as 157,735 in the most recent years.

Variable	Obs	Mean	Std. Dev.	Database	Min	Max	P1	P10	P25	P50	P75	P90	P99
	1,073,568	1908	140.305	Full sample	-2285	2015	1381	1820	1894	1943	1971	1985	1994
Birth year	139,965	1902	164.046	Wikipedia	-1570	2015	1220	1811	1893	1944	1974	1990	1997
	933,603	1909	136.369	Freebase	-2285	2015	1406	1821	1894	1943	1971	1985	1993
	499,959	1886	235.683	Full sample	-2566	2015	689	1712	1889	1954	1993	2008	2015
Death year	80,997	1771	382.134	Wikipedia	-1991	2015	254	1278	1730	1932	1985	2007	2015
	418,962	1908	187.17	Freebase	-2566	2015	1075	1790	1899	1958	1994	2008	2014

Table 7: Statistics on birth year and death year

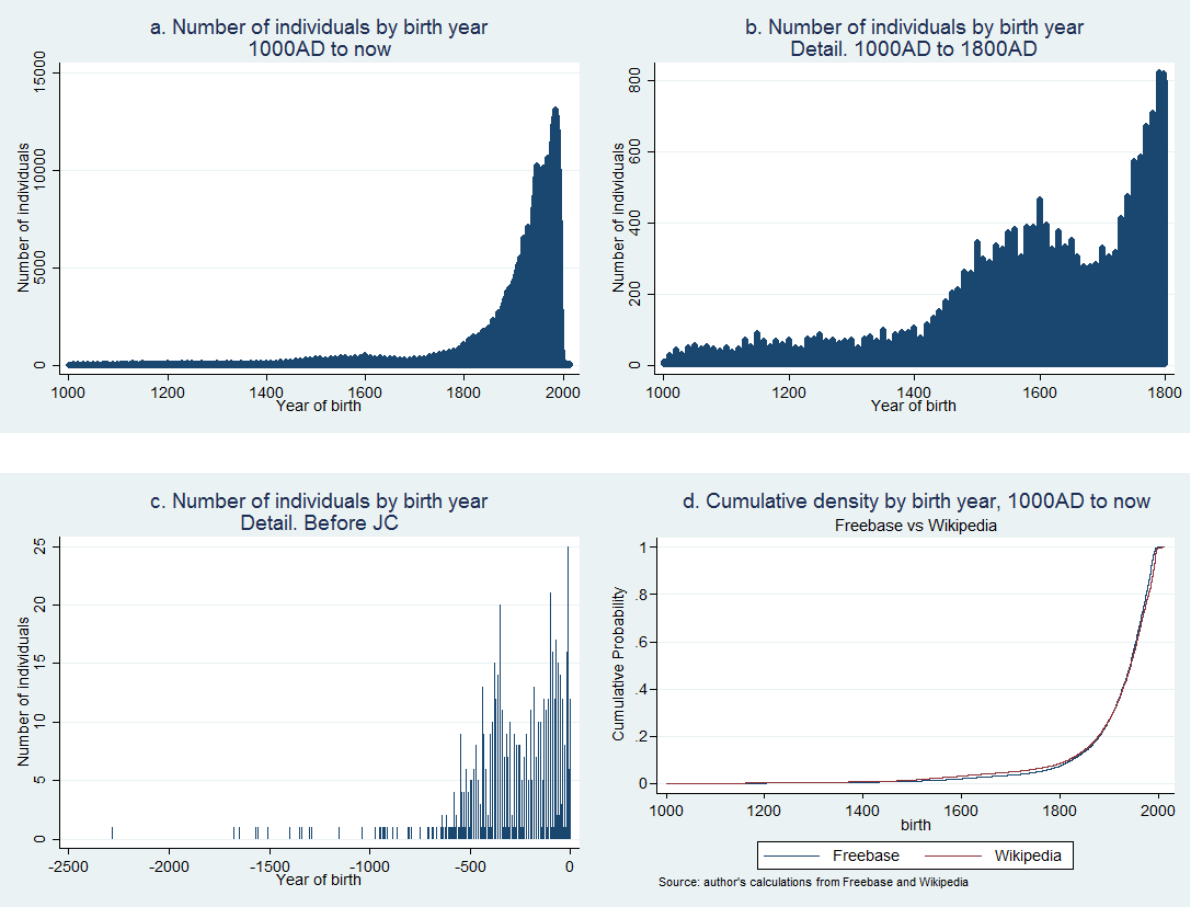
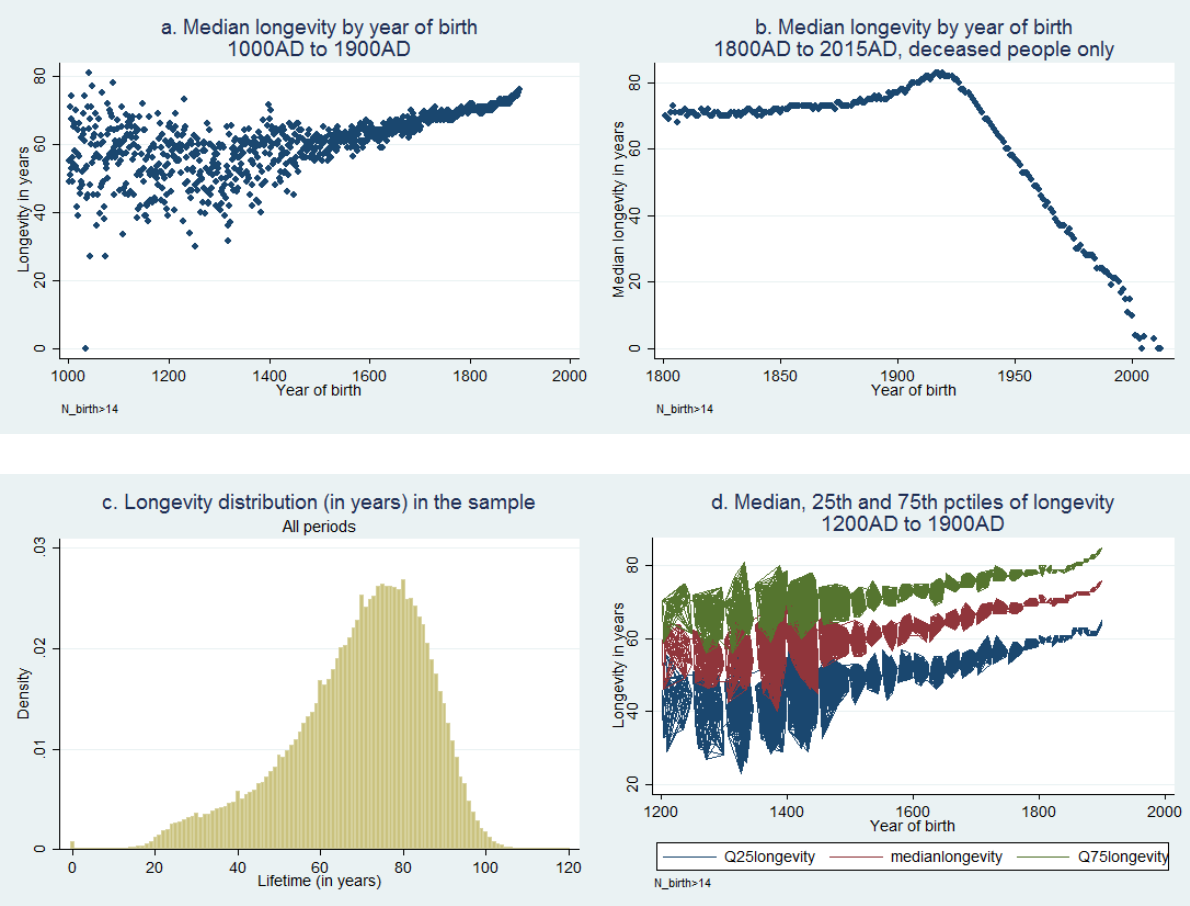


Figure 6: Cohort size in the database of individuals, `Freebase/Wikipedia` samples merged (first three charts) or kept separated bottom left chart

2.2 Average longevity

Longevity is expressed in years and computed as the difference between year deceased and year of birth, each of these itself estimated (see Section 1) from all available information on birth and death coming from `Wikipedia categories`, `the Infobox` and the `Abstract/Main text`. We first present the overall distribution for all years, then for the most recent period. For individuals born after 1900 and for

whom a death has been recorded, lifespan is necessarily declining over time. For individuals born in 2000, average lifespan must be necessarily computed using 2015 as a reference year along with his/her birth year. As found in de la Croix and Licandro (2015), we observe that the steady improvements in longevity start with cohorts born around 1600.



Note: After 1900, a majority of individuals is still alive and the series is unreported, since by construction life duration decreases to zero as years of birth are closer to 2015.

Figure 7: Longevity in years

Interestingly we see from Figure 7a and 7d that the lifespan variance across individuals also decreased over the sample period; although it rose again after 1830, presumably because of an increase in people of an older age.

2.3 Average visibility: all individuals

The large number of individuals in the database is a distinctive feature of our work, as compared to previous attempts. This large number necessarily implies that most individuals have little visibility on the web. Yet, they may contribute to the social, economic or cultural development of their area of residence.

We start by providing an overview of the distribution of the number of words and translations of the biography. All distributions, `Freebase` and `Wikipedia`, are very skewed. Pages that were collected using `Freebase` are on average longer and more translated than these extracted directly from `Wikipedia` following the "bottom-up" approach. Indeed, the additional pages from the `Wikipedia` search bring many individuals with little visibility. Another interesting finding is the time evolution of these two variables: as time goes on, they tend to decrease suggesting that history has kept memory of the most notable figures, while more recent individuals are more numerous and on average less impactful.

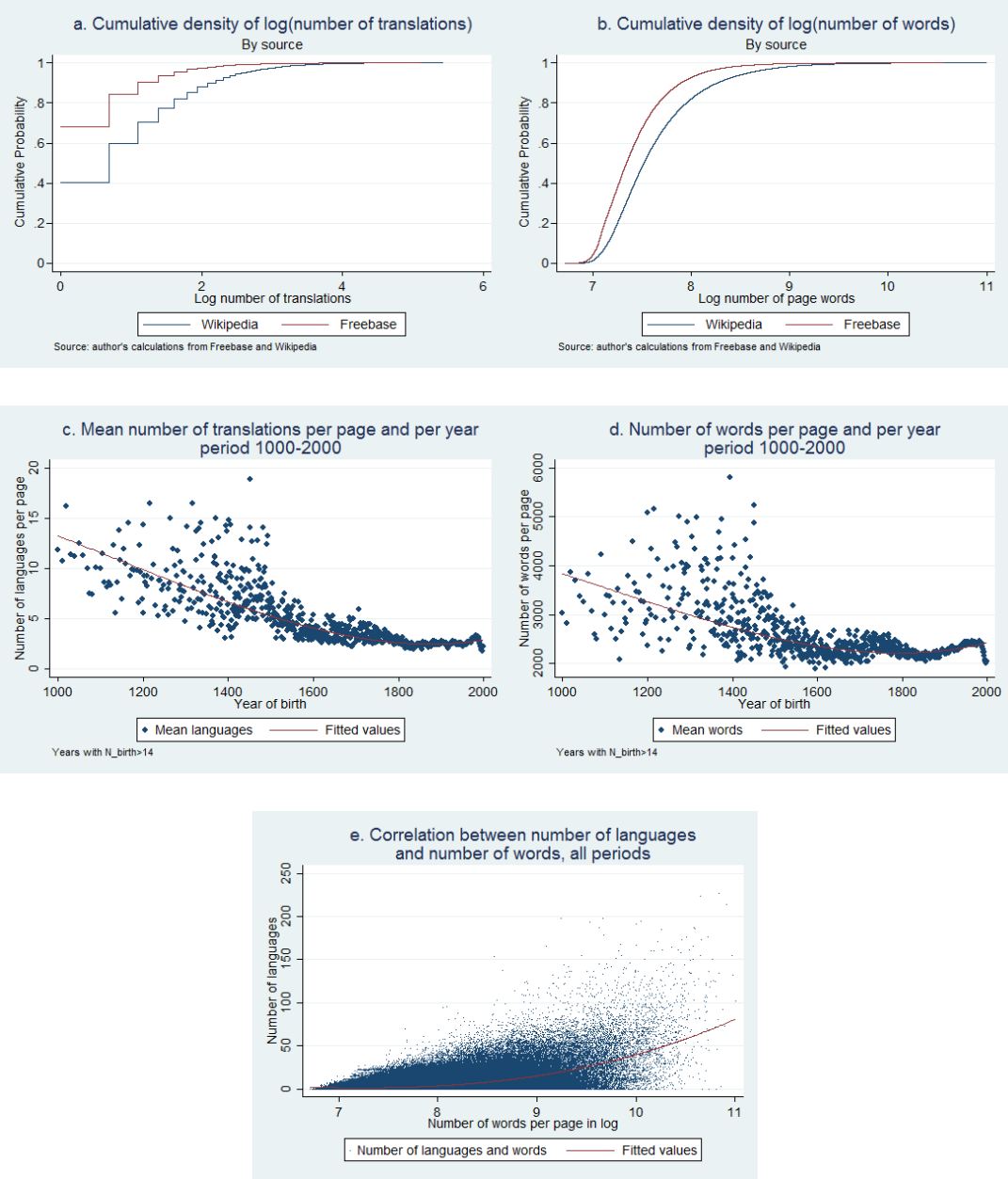


Figure 8: Components of visibility: distributions of number of translations of pages and number of words, time evolution and correlations

Both indicators (number of words and number of page translations) are correlated with many other variables, such as "number of bibliography items, number of footnotes, number of "See also", number of references, number of external links, number of "Further readings", or number of "Headlines" as indicated in Table 8.

These different components are used in what follows to create two possible visibility indices. It turns

Variable	Mean	Std. Dev.	Min	Max	P1	P5	P10	P25	P50	P75	P90	P95	P99
Languages	2.507	5.944	0	227	0	0	0	0	1	2	6	11	28
Number of words	2261.411	1764.577	826	60421	1060	1162	1234	1423	1768	2422	3606	4913	9455
Bibliography	.452	6.508	0	488	0	0	0	0	0	0	0	0	11
Footnotes	2.746	9.896	0	392	0	0	0	0	0	0	0	26	47
See also	.335	3.06	0	290	0	0	0	0	0	0	0	2	4
References	16.093	28.242	0	430	0	0	0	0	1	11	71	77	90
External links	9.931	24.032	0	275	0	0	0	0	0	1	65	71	82
Further readings	.126	2.501	0	426	0	0	0	0	0	0	0	0	1
Headlines	4.075	3.711	0	96	0	1	1	2	3	5	8	11	18

Table 8: Number of words per page, translations and other visibility indicators

out that the most basic combination leads to relatively sensible rankings. Indeed, the log of the product of the number of translations (+1) times the number of words in the `Wikipedia` biography leads to an indicator, called visibility, which is represented in Figure 9a. An alternative ranking is composed as the first index (visibility) multiplied by the log of the sum of all of the other variables described in the previous paragraph. The correlation between both indicators is 0.90 in levels and 0.88 in logs. The second distribution (Figure 9b) is more symmetrical than the one based on the basic visibility index positively skewed. The distribution in Figure 9 is based on visibility and will be used in the rest of the paper as it best reflects the skewness of visibilty in the database (already captured by the fact that we use logs; in levels, the distribution is extremely skewed leftward). It is also no surprise that visibility is higher when individuals are selected from `Freebase` given that they have shorter pages and fewer translations, as indicated above (Figure 9c); and when they belong to older cohorts (Figure 9d).

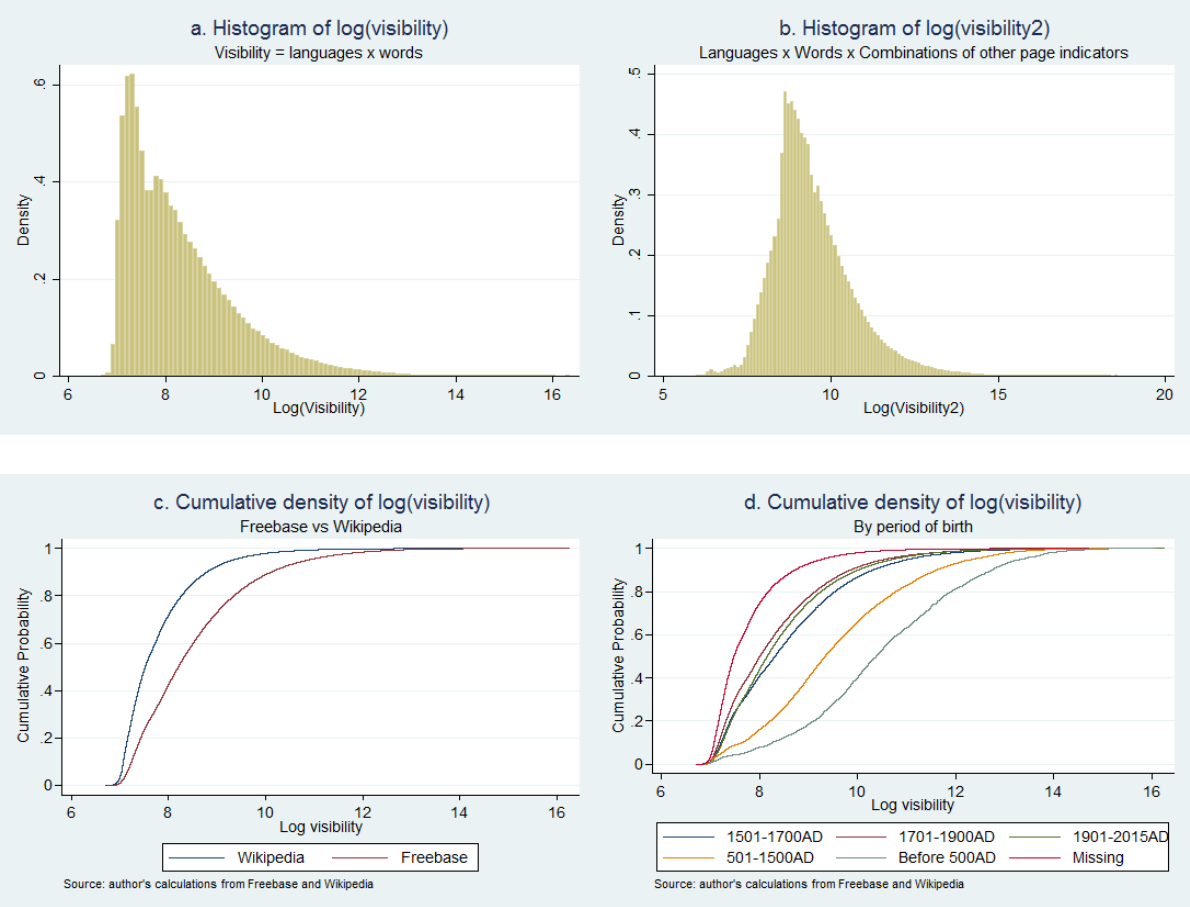


Figure 9: Visibility indices

2.4 Visibility: women

Over time, an interesting pattern emerges, a U-shape curve, with a local minimum at around 1700 (Figure 10a). This might be attributable to a composition effect (e.g. fewer and fewer individuals in the Family group and more and more in the Artists and Sports categories). However, unreported graphs, available from authors upon request, reveal that the U-shape pattern emerges for all six groups of occupations. At the end of the observation period, the female share is at around 0.25. We see also that females are less visible than males; there is a clear first order stochastic dominance in Figure 10b as expressed by the c.d.f. of visibility.

Occupation	Obs	Mean	Std. Dev.	Min	Max	P1	P10	P25	P50	P75	P90	P99
Entertainment	698,704	1937.053	85.996	-1557	2015	1610	1877	1921	1958	1979	1988	1995
Academics	130,075	1886.951	145.864	-1570	2015	1428	1803	1870	1921	1947	1963	1989
Entrepreneur	76,262	1879.818	108.782	-550	2015	1510	1777	1841	1907	1949	1966	1987
Family	21,338	1767.293	323.474	-1398	2015	191	1422	1737	1886	1945	1972	2000
Governance	303,263	1848.069	205.971	-2285	2015	862	1709	1833	1905	1946	1962	1986
Other	39,721	1893.812	151.736	-915	2015	1298	1814	1882	1922	1959	1979	1993

Table 9: Summary statistics: birth year by occupations

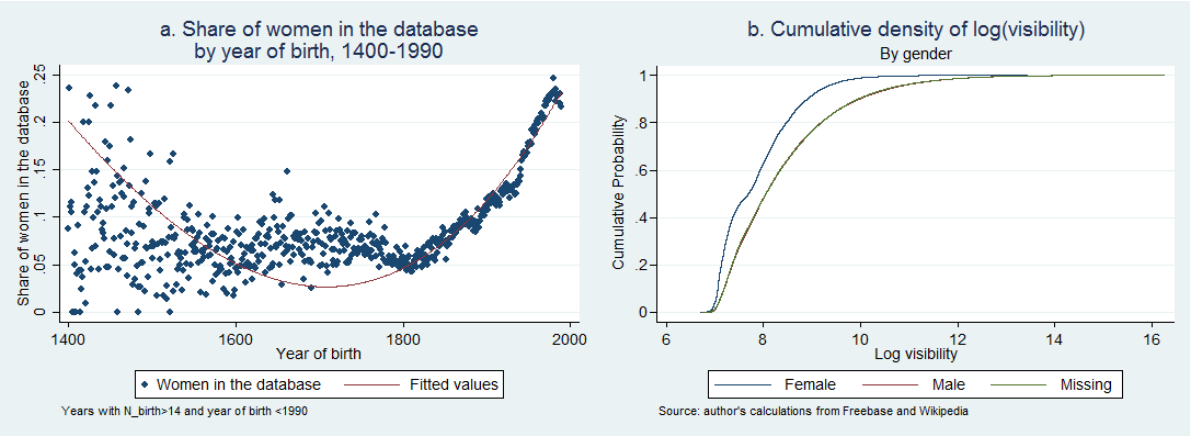


Figure 10: Share of women in different occupations throughout history

2.5 Evolution of occupations

We now investigate the share and visibility of our different occupations. As indicated, we have three levels of aggregation for occupations. Table 9 shows the distribution of the highest level of aggregation (six categories).

Figure 11a shows the evolution of these categories over time, for people born before 1990 (e.g. being at least 25 y.o.). The series sum up to slightly more than 1 since an individual may be in more than one category (e.g. Academics and Entrepreneur; the only exclusion is that an individual cannot be in Governance and another category). As visible, the post-1950 period sees the rise of the "Entertainment category", which by far dominates the database. The Governance category, most present until the beginning of the XIXth century, decreases after the 1840's cohort and drops further after 1950 (mechanically since these are shares, but gross numbers also decrease).

The middle part of the graph shows the evolution of categories for the intermediate level of aggregation (15 categories). In particular, it shows that Sports rose in two periods: for the birth cohorts between 1850 and 1870, with the emergence of sports contests in the second half of the XIXth century culminating with the first modern Olympic Games in 1896 and the first Tour de France in 1904. Interestingly, as Figure 11c shows, there has been a race between Sports and Art and Literature/Media between birth cohorts 1860 to 1950 with the final victory of Sports after 1950. The Figure 11b shows the cumulative

density of the visibility index retained; most activities are similar in terms of prominence except Family, which tends to be over-represented at large visibility levels.

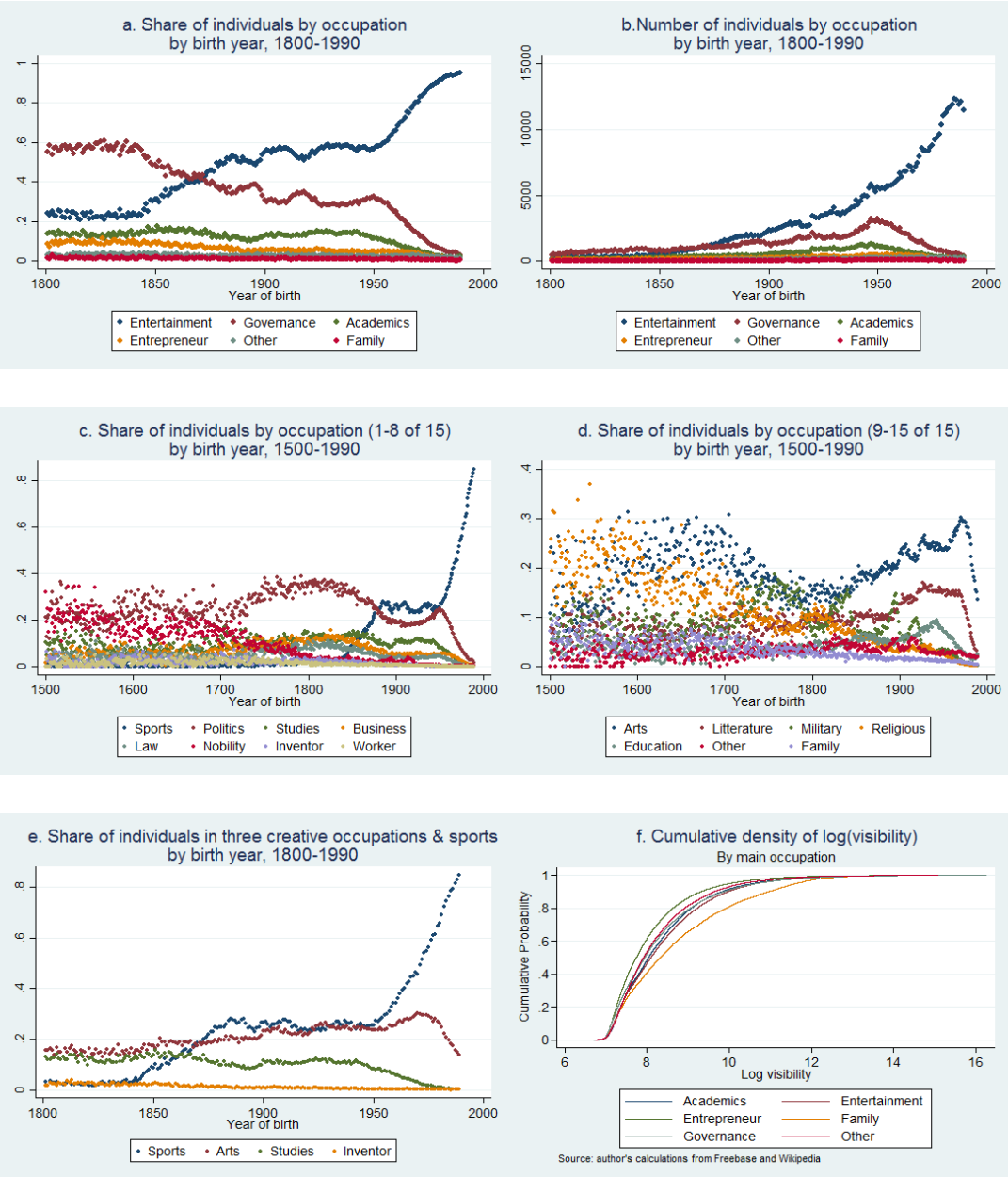


Figure 11: Share of occupations and creative occupations throughout history

2.6 Detailed lists of individuals in the database: top people and examples in various places

We now present the list of individuals in various categories. The top of each table lists the top 20 individuals in the category, and then sample individuals in the top decile, top quartile, median, third quartile and last decile, to give an overview of the composition of the database. All other categories are reported in Appendix.

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	3	Jesus	0	30	.	Family	religious
2	5	Napoleon Bonaparte	1769	1821	France	military	politics
3	6	Winston Churchill	1874	1965	England	politics	politics
4	9	Adolf Hitler	1889	1945	Austria	politics	politics
5	10	Joseph Stalin	1878	1953	Russia	politics	politics
6	12	Mahatma Gandhi	1869	1948	India	politics	politics
7	13	Mahomet	570	632	.	religious	Other
8	14	William Shakespeare	1564	1616	England	lit	lit
9	18	Alexander The Great	-356	-323	Greece	nobility	nobility
10	19	Mustafa Kemal Atatürk	1881	1938	Turkey	military	military
11	20	Franklin D. Roosevelt	1882	1945	US	politics	politics
12	21	Abraham Lincoln	1809	1865	US	politics	politics
13	23	George Washington	1732	1799	US	politics	military
14	26	Charlie Chaplin	1889	1977	England	arts	arts
15	27	Vladimir Lenin	1870	1924	Russia	politics	politics
16	28	Charles de Gaulle	1890	1970	France	lit	politics
17	30	Albert Einstein	1879	1955	Germany	studies	studies
18	32	Karl Marx	1818	1883	Germany	studies	studies
19	41	Theodore Roosevelt	1858	1919	US	politics	lit
20	43	Vincent Van Gogh	1853	1890	Netherlands	arts	arts
		Top decile (random sample)					
25019	101995	Karl Freiherr Von Müffling	1775	1851	Germany	military	
25021	102007	Custodio García Rovira	1780	1816	Spain	politics	arts
25020	102008	Jacques de Billy	1602	1679	France	religious	studies
25023	102023	Antoine-Vincent Arnault	1766	1834	France	arts	arts
25022	102026	Ebbo	775	851	Germany	religious	religious
		First quartile (random sample)					
62554	268965	Charles L. Hutchinson	1854	1924	US	business	business
62553	268967	Emil Johann Lambert Heinricher	1856	1934	Austria	studies	studies
62550	268969	George Douglas, 16th Earl of Morton	1761	1827	US	Family	nobility
62552	268977	Paul Morgan (actor)	1886	1938	Austria	arts	arts
62551	268983	Emil Barth	1879	1941	Germany	politics	worker
		Median (random sample)					
125104	591207	Edmund Frederick Erk	1872	1953	US	politics	politics
125102	591276	Ripley Hitchcock	1857	1918	US	lit	arts
125106	591293	Alexis Lesieur Desaulniers	1837	1918	Canada	law	politics
125105	591336	Anna Maria Hilfeling	1713	1783	Sweden	arts	arts
125103	591347	Edward Garrard Marsh	1783	1862	England	lit	religious
		Third quartile (random sample)					
187655	931626	William Cole (scholar)	1753	1806	England	education	education
187656	931829	A. W. Andrews	1868	1959	England	studies	lit
187658	931846	Thomas Nash (Newfoundland)	1765	1810	Ireland	worker	Other
187657	931911	Hugh Aiken Bayne	1870	1954	US	Family	law
187654	931912	Samuel Backhouse	1554	1626	England	business	politics
		Last decile (random sample)					
225187	1121555	Edward Hopkins (politician)	1675	1736	Ireland	politics	education
225185	1121647	Braxton Lloyd	1886	1947	US	studies	politics
225188	1121709	Stephen Furniss	1875	1952	Canada	politics	politics
225189	1121712	Ida Holterhoff Holloway	1865	1950	US	arts	arts
225186	1121786	Robert Henry Blosset	1776	1823	England	law	military

Table 10: Individuals born before 1891

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	1	Barack Obama	1961		US	politics	education
2	2	Ronald Reagan	1911	2004	US	politics	arts
3	4	George W. Bush	1946		US	politics	business
4	7	Nelson Mandela	1918	2013	South Africa	politics	politics
5	8	Michael Jackson	1958	2009	US	arts	arts
6	11	John F. Kennedy	1917	1963	US	politics	politics
7	15	Pope Francis	1936		Argentina	religious	religious
8	16	Cristiano Ronaldo	1985		Portugal	sports	sports
9	17	Pope John Paul Ii	1920	2005	Poland	politics	religious
10	22	Roger Federer	1981		Switzerland	sports	sports
11	24	Lionel Messi	1987		Argentina	sports	sports
12	25	Novak Djokovic	1987		Serbia	sports	sports
13	29	Hillary Rodham Clinton	1947		US	politics	politics
14	31	Bill Clinton	1946		US	politics	politics
15	33	Pope Benedict Xvi	1927		Germany	politics	religious
16	34	Che Guevara	1928	1967	Argentina	politics	politics
17	35	Elvis Presley	1935	1977	US	arts	arts
18	36	Mao Zedong	1893	1976	China	politics	politics
19	37	Hugo Chávez	1954	2013	Venezuela	politics	politics
20	38	Rafael Nadal	1986		Spain	sports	sports
			Examples at the top decile				
81954	113355	Evan Jenkins (politician)	1960		US	politics	politics
81956	113358	Anton Golotsutskov	1985		Russia	sports	sports
81957	113361	Alphonse Leweck	1981		Luxembourg	sports	sports
81953	113365	Lazar Ristovski	1952		Serbia	arts	arts
			Examples at the first quartile				
204889	283176	José Greci	1941		Italy	lit	arts
204888	283204	Valeria Solarino	1979		Italy	arts	arts
204887	283206	Mark Preston	1968		Australia	business	business
204886	283212	Lev Dobriansky	1918	2008	.	education	education
			Examples at the Median				
409775	567684	Jeanette Lunde	1972		Norway	sports	sports
409774	567752	Cristian Andrés Campozano	1985		Argentina	sports	sports
409776	567799	Clifford Peeples	1970		Ireland	religious	politics
409773	567823	Stephen Adams (business)	1937		US	business	business
			Examples at the third quartile				
614663	875000	Dorice Reid (baseball)	1929		US	sports	sports
614660	875029	Birger Wernerfelt	1951		Denmark	studies	business
614661	875091	Les Phillips	1963		England	sports	sports
614662	875145	Casey Henwood	1980		New Zealand	sports	sports
			Examples at the last decile				
737592	1080475	Philip Gardiner	1946		Australia	politics	politics
737595	1080596	James A. Andersen	1924		US	politics	law
737596	1080687	Jeanette Kuvin Oren	1961		US	arts	arts
737594	1080777	Bruce Martyn	1930		US	sports	lit
737593	1080969	Claude Legris	1956		Canada	sports	sports

Table 11: Individuals born after 1891

2.7 Geographical density through history

Finally, the sample is predominantly drawn from the Western World (Europe and North America) with, however, a rise of other continents, mostly Asia and Latin America over the sample period (see Figure 12a). The three main European countries have a share that declines over the centuries (Figure 12b). Also note the over-representation of the United Kingdom as compared to France and Germany in the database. Interestingly, Figure 12c shows that individuals from less represented countries tend to have higher visibility, due to selection of the sample.

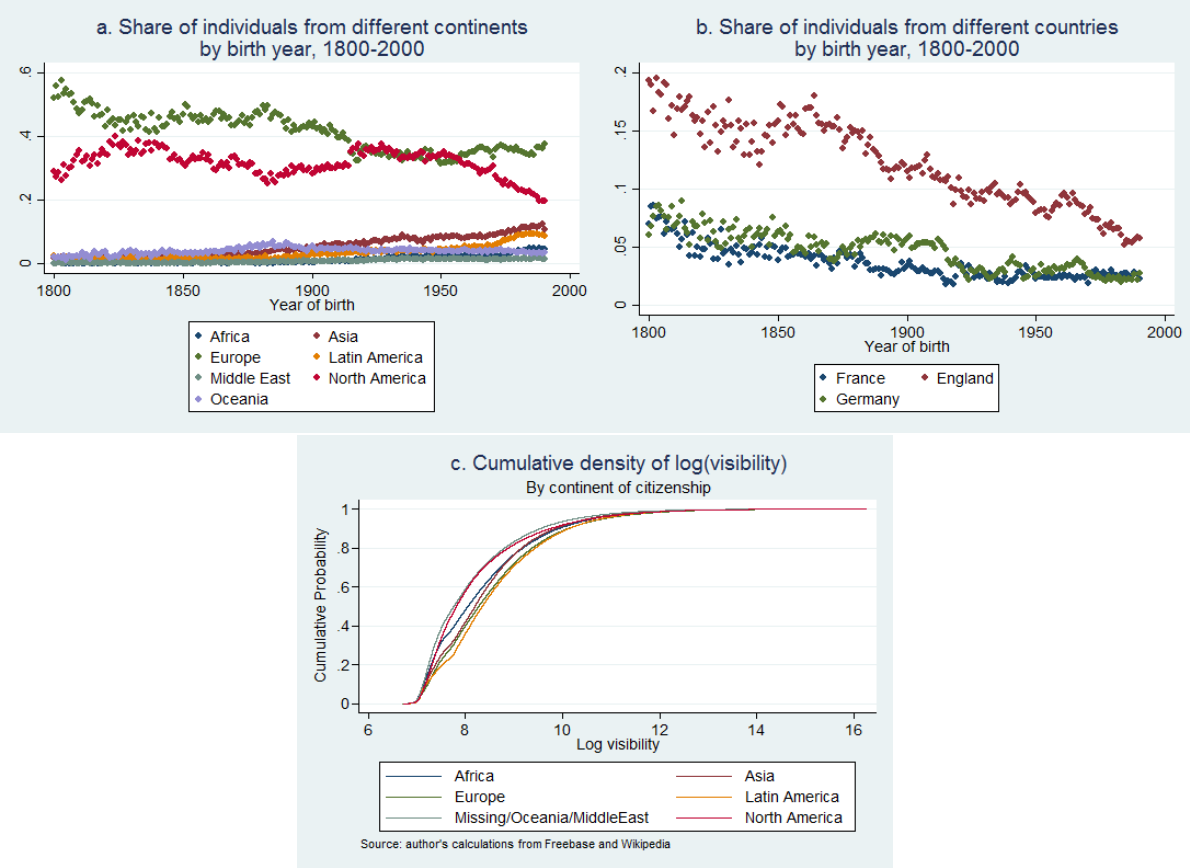


Figure 12: Geographical composition of the sample

3 Facts on geographical mobility

3.1 Between birth and death

We have 842,356 individuals with geocoded places of birth, and 264,572 individuals with a geocoded place of death. Note that some of them have a birth year or a death year missing. Overall, we obtain geocoded information on both places of birth and death for 227,223 individuals. To check whether the final sample is biased toward more prominent people, we compare the distributions of visibility across samples.

In the latter sample, the median distance between birth and death is 268 kilometers, the mean is 1522km. Various percentiles are represented in Table 12. There is no systematic difference across visibility levels, and the very top end of the visibility distribution is even associated with lower distances between locations of birth and death. With regards to secular evolutions, one obtains the minimum of distances from the individuals born in the middle-ages, with those born before the 6th century having slightly higher distances (this may be a composition effect rather than a trend affecting the overall population born in these periods). After 1500AD, distances start increasing again with a particularly large increase in median distances and a larger increase in top 10% distances due to the existence of settlements in the new continents.

Period	Obs	Mean	Std. Dev.	Min	Max	P1	P25	P50	P75	P99
Before 500AD	3066	676.468	796.802	0	3212.737	0	3.721	316.725	1125.775	3212.737
501-1500AD	32522	505.117	1432.714	0	14715.14	0	0	100.131	389.473	8857.269
1501-1700AD	49669	926.074	2113.387	0	16684.94	0	22.546	156.78	515.041	9988.019
1701-1900AD	707193	1650.343	3252.779	0	19821.38	0	56.313	318.805	1334.502	16970.39
1901-2015	640993	1729.322	3059.098	0	19852.35	0	54.353	368.549	1808.116	14567.91
Missing	27952	1561.39	3139.575	0	19805.96	0	.729	189.147	1118.294	15859.45

Table 12: Distance from birth to death, in kilometers, all sample and by periods of history

Period	Obs	Mean	Std. Dev.	Min	Max	P1	P25	P50	P75	P99
Before 500AD	10,560	16.517	17.283	0	101	0	5	10	23	75
501-1500AD	79,229	16.156	20.182	0	161	0	5	10	19	110
1501-1700AD	135,133	11.674	18.754	0	277	0	4	7	12	78
1701-1900AD	1,281,826	10.872	12.377	0	218	0	5	8	12	64
1901-2015AD	3,087,414	14.403	30.951	0	534	0	3	7	13	158
Missing	439,980	7.534	17.147	0	344	0	2	4	8	58

Table 13: Number of `GeoLinks` per individual (Full sample and by periods)

3.2 Summary Statistics

Based on the methodology explained above in Section 1, we can compute distances bewteen birthplace and any `GeoLink`, and provide summary statistics.

Visibility Perc.	Obs	Mean	Std. Dev.	Min	Max	P1	P25	P50	P75	P99
All	3,987,369	8.942	16.467	0	479	0	3	6	10	66
90	368,459	19.581	35.094	0	355	0	5	9	18	196
95	449,297	29.623	51.754	0	534	1	8	14	28	277
99	190,854	35.729	38.114	0	306	4	14	24	43	194
99.9	38,163	45.877	28.566	0	158	7	25	39	59	142

Table 14: Number of `GeoLinks` per individual (most visible individuals)

Period	Obs	Mean	Std. Dev.	Min	Max	P1	P25	P50	P75	P99
Before 500AD	4,732	1029.195	1654.817	0	13248.77	0	46.953	497.799	1284.738	9102.902
501-1500AD	45,551	729.829	1696.815	0	18083.47	0	31.204	197.733	595.974	9121.377
1501-1700AD	83,482	951.713	2183.658	0	19581.95	0	11.268	150.668	557.502	10522.65
1701-1900AD	1,007,984	1618.402	3312.81	0	19826.06	0	9.897	246.536	1179.298	16908.33
1901-2015	2,542,850	1805.327	3365.142	0	20012.04	0	1.816	259.298	1671.164	15985.91
Missing	144,790	1837.369	3468.855	0	19843.5	0	.021	179.772	1621.88	15745.76

Table 15: Distance from birth to any identified `GeoLink` per individual (in km, Full sample, by periods)

Visib. Perc.	Obs	Mean	Std. Dev.	Min	Max	P1	P25	P50	P75	P99
All	2,903,945	1559.643	3199.134	0	20,012.04	0	.092	176.389	1207.587	16,287.4
90	313,248	1970.395	3459.575	0	19,789.02	0	19.559	389.453	1874.146	16,365.76
95	401,799	2286.674	3681.198	0	19,941.56	0	67.235	549.752	2549.724	16640.06
99	175,629	2567.837	3781.461	0	19,936.87	0	145.157	740.887	3337.66	16453.98
99.9	34,768	2573.893	3672.483	0	19,243.99	0	171.841	785.636	3473.262	15,716.43

Table 16: Distance from birth to any identified `GeoLink` in individual's life, in km (most visible individuals)

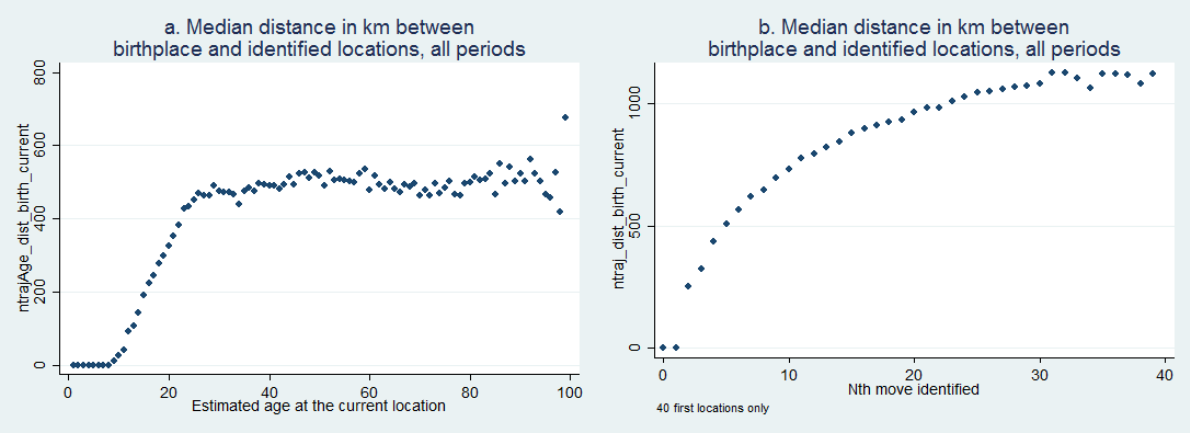


Figure 13: Average distance between birth place and current location

It can be seen that the typical individual leaves his or her birthplace between 10 and 20, and after 20 the median distance of the individual to his or her birthplace is over 200 km.

Trajectories	Origin	Destination	#	Mean Percentile Visib.
1	UK	USA	35,061	64.5%
2	CAN	USA	28,202	60.0%
3	GER	USA	15,541	66.5%
4	US	CAN	13,816	58.0%
5	UK	AUS	13,719	49.6%
6	UK	FRA	9,702	67.7%
7	IRE	UK	9,636	61.1%
8	USA	FRA	8,005	69.1%
9	AUS	UK	7,706	64.4%
10	ITA	US	7,098	71.5%

Table 17: Most common country-to-country trajectories

3.3 Identified international mobility

One recovers the full addresses, including the country information of all geocodes in the database (current location, places of birth and death when available). The country of origin corresponds to borders in 2015. We round up after the third digit the geocoordinates of all `GeoLinks` including places of birth and death. We obtain a grand total of 276,677 unique locations identified in the database. Of those, we match a total of 259,109 locations with a country and most of the time a full address, using the command geocode3 and a specific application, `Google Maps Geocoding API`. We then match these countries to the original locations. Of all non-unique geolinks present in the database, one is able to match 4,859,007 with a country. As for the individual's database, the U.S. represents the largest share (1,831,479 lines), followed by the UK (821,807 lines), then Germany, Canada, France, Italy, etc. The total number of countries, including overseas territories such as Mayotte or Saint-Pierre et Miquelon, islands such as Turks and Caicos Islands, U.S. Minor Outlands, Guam etc. is 248. Only 35,421 geocodes are returned as "not found" by the `Google App`. We finally create a variable "move_country" if the country of birth differs from the country of the current geolink, and none of the birth and current countries are missing or "not found". In the database, 19.7% of identified geolinks with a country have a country of birth different from the country of the individual's current location.

In time, we obtain a U-shape with a positive and accelerating trend from 1960 to 2015, as Figure 14a shows. The right panel, however, shows that this is more due to a composition effect, the "Entertainment" category being more internationally mobile and having a growing share in the sample as time goes on. Nevertheless, all categories exhibit a rising fraction of country-to-country mobility.

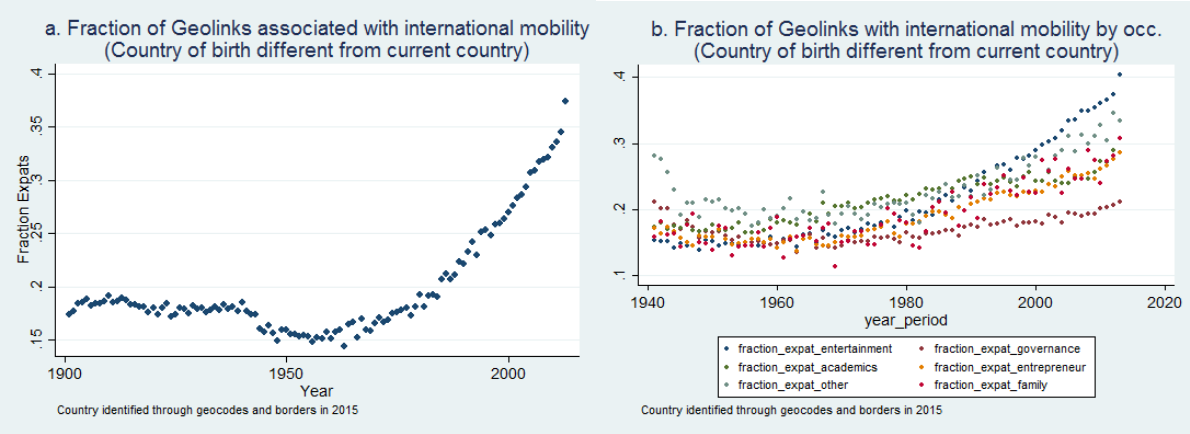


Figure 14: Country to country mobility

4 Connecting people to cities

In this section, we introduce a new unified database with global historical urban population. We have different data sources covering distinct periods but with different population concepts: some cover urban areas, some other cities defined by administrative boundaries. Most Censuses date back to the beginning of the 19th Century and indeed cover city population but not agglomeration population. To our knowledge no unified database with Census historical data has yet been gathered. We present first the different Census sources identified and collected through scraping and OCR and compile a global historical Census population database.

We then complete the population database with available data from other sources for different periods: 1500-1800 (Urbanisation Hub-Bairoch-Bosker) which covers cities and the surrounding areas when contiguous, then 1800-2010 (Lincoln Institute of Land Policy) for the largest cities outside Europe, and finally after 1950 (UN Population database). Figure 1 below illustrates how these different sources, either official or academically certified, complement each other.

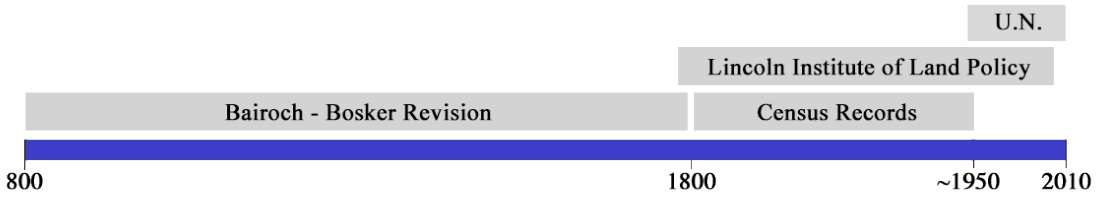


Figure 15: Timeline of sources used

After the urban population data was collected, city names were geocoded using the `Google Maps API` and matched with the trajectories of `notable people`.

4.1 Data collection

4.1.1 Period 1800-2010: Census records and Lincoln Institute

We collected and assembled detailed and official Census records. Census collections started in the early-to-mid-19th century. Census records are available primarily from National Statistical Institutes and cover a larger number of 'cities': 36,541 communes in France, 9,148 communes in Italy, 8,915 urban centers in Canada, and 1,008 incorporated cities in the United States, etc.
In order to obtain historical census data, multiple methods of data extraction were used. When downloadable digitized census records were not available, web scraping (for 6 countries) and optical character recognition (OCR for 3 countries) helped extract the required data. Certain manipulations on the extracted data were required, such as aggregating city data to urban areas, matching city names across datasets, and combining male/female portions of urban populations. More information on these different methods, but also on official links used to access the data, city definitions, dates, number of observations and years available and more are reported in Table 18. .

In a few instances, city limits and definitions changed over time, due to land organization differences, evolving urban administrative regions, etc. An example is the UK where data are consistent between 1801 and 1911, but inconsistent with the period 1921-1961. In the longitudinal analyis, we treat these two samples a set of distinct cities (e.g. an unbalanced panel) to avoir dealing with non-comparable datasets. A list including all manipulations performed in the data extraction process is included in the Online appendix. Official census records have been collected for 2 countries in Asia (India and Japan), 11 countries in Europe (Austria, Belgium, Denmark, France, Germany, Italy, Netherlands, Portugal, Russia, Switzerland, United Kingdom), 2 countries in America (Canada and United States) and 2 countries in Oceania (Australia, New Zealand). Information for 65,087 different cities around the world has been gathered. Countries have been chosen both for the primary role they played over the course of history and the availability of official census records for these countries.
To complement the Census dataset over the same period, we use a database compiled by the Lincoln Institute of Land Policy which provides historical urban population data for 30 major urban agglomerations from 1794 to 2005. The data is reported in intervals of 20 or 25 years and was used to fill gaps of missing data where city limits permitted. This database comprises the following cities sorted by continent and sub-region: Africa (Accra, Algiers, Cairo, Johannesburg, Lagos, Nairobi), Asia (Bangkok, Beijing, Kolkata, Mumbai, Shanghai, Tehran, Tel Aviv), Central America (Guatemala City, Mexico City), Eurasia (Istanbul, Moscow), Europe (Warsaw), Middle East (Jeddah, Kuwait City), South America (Buenos Aires, Santiago, Sao Paulo) and Southeast Asia (Manila). We only used information concerning cities for which we failed to find reliable census data and thus did not consider Lincoln data for Chicago, Los Angeles, Paris, London, Sydney, or Tokyo.
Using Census years (typically every 10 years), we linearly interpolate population for each city in the databse to obtain yearly population and then collapse city population into periods of time of 10 years intervals (e.g. 1801-1810, 1811-1821 etc to 2001-2010).

4.1.2 Backward extension 800AD–1800AD: The Urbanisation Hub-Bairoch-Bosker database

To account for city population from 800 to 1800, we used an urban population database compiled initially by Bairoch et al. (1988) and further revised by Bosker et al. (2013). The maximum number of

Continent	Country	City Definition	Dates	Number of years	Max # of Obs.	Source	Data Extraction Method
Asia	India	Towns and Suburbs or Cantonments	1863-1911	7	272	English Parliament / University of Chicago	Excel download
	Japan	Shi	1873-2010	24	50	Statistics Japan	Excel download
Europe	Austria	Communes	1869-2001	14	200	`Wikipedia` (cross-referenced with Statistik Austria)	Web scrap
	Belgium	Communes	1846-2015	5	571	Statistics Belgium	OCR
	Denmark	Urban Areas & Islands	1769-2015	72	20	Statistik Banken	Excel download
	France	Communes	1793-2006	34	36,541	Cassini	Web scrap
	Germany	Großstadt	1816-2013	31	76	`Wikipedia` & Deutsche Verwaltungsgeschichte	Web scrap
	Italy	Communes	1842-2011	16	9,148	Stat. Bureau of Italy	Web scrap
	Netherlands	Village/Municipalities	1795-1919	10	2903	Stat. Bureau of the Netherlands	Excel download
	Portugal	Localidades com mais de 10 milhares habitantes	1864-2011	15	310	National Statistical Institute of Portugal	Excel download
	Russia	Urban Center	1750-2001	38	163	Populstat	Web Scrap
	Switzerland	Communes	1850-1990	15	3019	Stat. Bureau of Switzerland	OCR
	United Kingdom	Pre-1841: Urban Centers Post-1841: Urban Areas	1801-1911	12	934	UK Data Archive / Cambridge U. / UK Census	Excel download
		Municipal Boroughs	1921-1961	4	381	Vision of Britain / United Kingdom Census	Web scrap
North America	Canada	Urban Centers	1871-1961	11	1,008	Statistics Canada	OCR
	United States	Incorporated Cities	1790-2010	23	8,915	US Census Bureau & Populstat	Excel download
Oceania	Australia	Capital cities (8)	1901-2011	96	8	Statistical Bureau of Australia	Excel download
	New Zealand	Urban agglomerations	1916-2006	11	568	Statistical Bureau of New Zealand	Excel download

Table 18: Census data: sources and main characteristics

European cities in the revised version of the database amounts to 677 cities in the year 1800. All cities in Bosker at al. (2013) include more than 10,000 inhabitants and are generally defined as urban areas, where "suburbs (faubourgs) surrounding the center" are included. Bosker et al. updated Bairoch's database by scanning recent literature concerning the major cities covered. In particular, they updated all cities which during a point in history were larger than 60,000 inhabitants. This led to a number of important revisions of population records concerning Muslim cities in medieval Spain but also for the cities of Palermo, Paris, Bruges, and London. Although these revisions decrease the total number of cities covered, the remaining observations are thought to be more reliable.

Bosker et al. (2013) additionally added Middle Eastern and North African cities to Bairoch et al. (1988), which previously only included European cities. This allowed us to obtain historical urban population data on a total of 116 cities coming primarily from the countries of Egypt, Turkey, the former Yugoslavia, Saudi Arabia, Oman, Yemen, Israel, Iraq, Lebanon, Libya, Tunisia, Algeria, and Morocco.[4]

Using available years (typically every 50 years or 100 years before 1000AD), we linearly interpolate population for each city in the databse to obtain yearly population and then collapse city population into our `Time Periods` of length of 50 or 25 years intervals (801-850 to 1651-1700, then 1701-1725, 1726-1750, etc.).

4.1.3 Forward extension 1950AD-2010AD: The United Nations database

A database provided by the United Nations' Department of Economic and Social Affairs gives urban agglomeration data from 1950 to 2010 at five year intervals. The definition of an urban agglomeration in this source references an urban area with over 300,000 inhabitants. This database was very useful since the data available from census records decreased steadily starting in the mid-20th century, as seen in Figure16 b below. The other advantage of this dataset compared to any other source of information is its worldwide coverage. In all, this database contains population data for 1,692 cities. We collapse the data into ten year intervals: 1951-1960, etc.

Keeping only the largest 50 cities in coutries larger than 300,000 sq. km and the largest 30 cities in other countries (or less if fewer cities where avaialble), we end up with a panel of cities linearly increasing from 800AD to 1800AD (Source: Bairoch-Bosker) and increasing by steps (as larger countries enter the database with a first Census) from 1800 to 1950. See Figure 16b. After 1950, the Census database decreases since the last Census available online varies from country to country (e.g. in the United Kingdom, the database terminates in 1961). We use after 1950 the UN database and its 1,692 cities.

[4]More details about this revision are available in the data appendix accessible from http://bit.ly/2029bhk

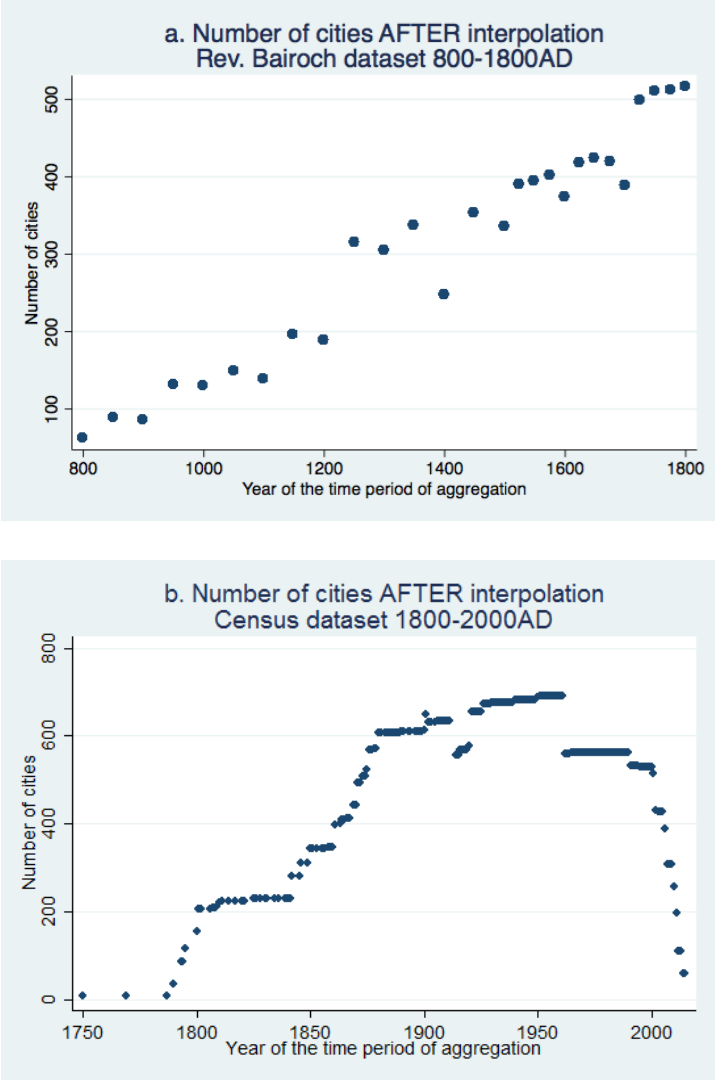


Figure 16: Number of cities by source of data (UH-Bairoch-Bosker ; Census), only retaining the largest 30/50 cities in each country

4.2 Distances to a large city ; at birth, at death, and in between

From now on, the city database is limited to the top 50 or 30 cities (for which population is available) in each country. We match the `GeoLinks` to the three nearest cities of this database period by period and only when population is available in that period. We use the `geonear` command, from geocodes of individuals' locations and of cities. It also returns distances in kilometers based on geodetic distances, "using a mathematical model of the earth" as specified in the description of the command. Figure 17 shows on the left hand side the cumulative density of distances of the third three cities. The Bairoch database covers well the individual's locations: over the period 800-1800AD, 60 percents of `GeoLinks` are located within a radius of 50 kilometers around a city in the database. Another 30% of `GeoLinks`

are within 50 kilometers of a second nearest city, and 17% are within 50 kilometers of the third nearest city. Similarly, 50% of individuals over the period 1800-1939 are within 50 kilometers of the closest city. Finally, the last period of the sample, based on the UN database, is the best matched: 80% of `GeoLinks` are within 50 kilometers of the closest city. On the right part of the graph, we represent the cumulative distance to the closest city by occupation. They do not differ much for the Census and the UN database of cities, but differ according to occupation over the period 800-1800, Academics and Entertainment (Arts, Literature and Media only over this period) being significantly more likely to be close to a big city.

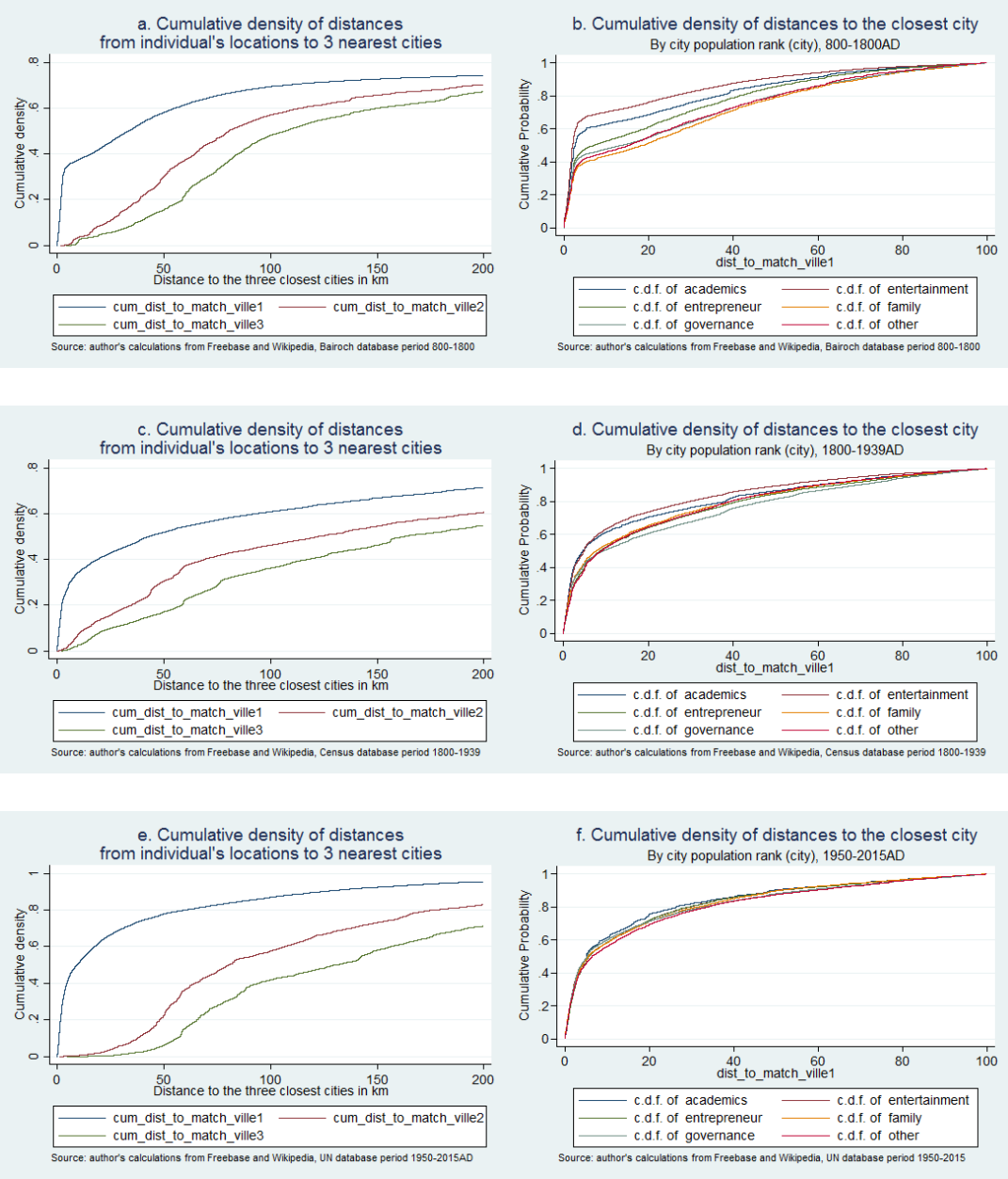


Figure 17: Distances between individual's locations and nearest cities in the three samples: Bairoch-Bosker ; Census ; United Nation Database

4.3 Facts on city rank, visibility and occupations

The last observations lead to a study of the relations between city size (or rank) and occupation and visibility. It can be seen from Figure 18 (left panels) that there is no systematic correlation between city rank in a country and the degree of visibility (in logs). For the top four cities, it turns out that the c.d.f. of log visibility are quite close to each other and the lowest c.d.f. is that of the first city and

this is true over the three sub-samples (800-1800AD, 1800-1939AD and post 1950). For instance, in the sample of Census data (1800-1939), the correlation coefficient between log visibility and city ranks from 1 to 50 is -0.08 and -0.07 if limited to the first 4 cities.
When the rank of the city is calculated for all countries, as in the right part of the Figure, things are different (e.g. over the period 1800-1939AD London, New York, Paris, Beijing are ranked as first cities in the world, etc.), the pattern is different. Individuals in the first city have higher visibility, and have lower visibility in the second, fourth, and then third cities in the world. The correlation coefficient between that rank from 1 to 4 is now positive, equal to 0.17. However, the correlation between city rank and visibility of individuals is negative again in the post 1950 period where the biggest cities in the world are in located in developing countries (bottom right chart).

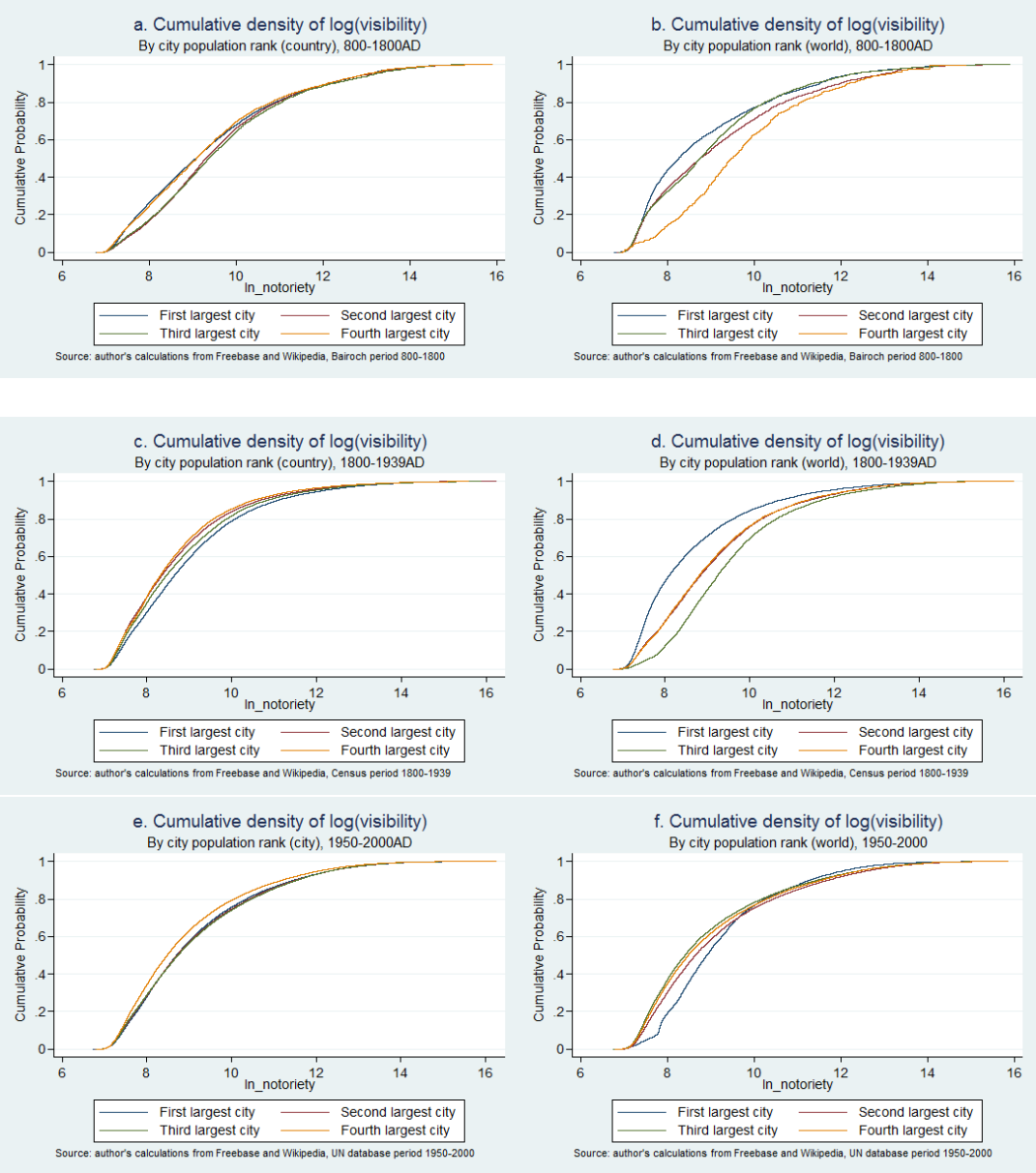


Figure 18: Links between city rank and visibility. Left charts: ranking within countries ; right charts: wolrd ranking. Top: Bairoch city dataset (800-1800AD); Middle: Census city datasets (1800-1939AD) ; bottom : UN city dataset (1950-2000AD)

Weights used	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$Visibility = transl. \times words$
Entertainment	65.3	79.0	117.1	3.1×10^8
Governance	50.3	59.6	87.3	5.2×10^8
Academia	22.2	27.8	41.4	9.3×10^8
Entrepreneurs	10.4	11.4	15.9	3.4×10^8
Family	2.1	2.8	4.3	1.2×10^8
Others	4.1	4.9	7.3	1.1×10^7
Total	154.3	185.4	273.1	9.7×10^8
# of cells: cities x periods	19,729	19,729	19,729	19,729

Table 19: Number of `notable people` by city/period weighted by the distance (periods 16-63 or 775AD-2005AD)

4.4 Creating city averages of the share of `notable people` across occupations

The next stage is to collapse the `GeoLinks` database into cities for each period analyzed. We do so for all locations for which the estimated age of the individuals is above 15 years. Instead of using an arbitrary distance threshold for the potential influence of an individual on a city, we use a continuous measure. The decay factor is $e^{-dist/33}$ where distance is measured in kilometers. This decay factor implies that only individuals exactly located in the center of the city have a full influence. At 10 kilometers from the centroid of the city, their influence is 0.73. At 20 kilometers, it is 0.50; at 50 kilometers, it is 0.38, etc. The procedure is replicated successively for the closest, second closest and third closest city: an individual placed at 20 kilometers of two cities presumably influences that city twice, but with weight 0.50.

We can also weight individuals according to their visibility. We use four different weighting procedures. The first one is to give equal weight to all individuals. The second one is to weight individuals by their percentile (in %) multiplied by 2 and divided by 100 (the median individual therefore has weight 1, the top person a weight of 2 and the bottom person a weight of 0). A third one is the square of the previous weight. The last one directly uses the visibility measures. That measure is very skewed and ranges from approximately 8000 to 10^7.

On average, over the period 775-2005, we represent in Table 19 the following weighted number of `notable people`.

A specific focus on the Entertainement category (Arts, Literature/Media, as well as Sports after 1850) is interesting. We calculate in Table 20 the shares of "efficient" units of `notable people` according to the population rank within the country. The different weights do not change much the pattern, except when weights are linear in visibilty. It also appears that the fraction continuously increases over time, from one fifth over the period 800-1800 to one third over the period 1800-1939 and finally one half after 1950. Figure 19 shows that the correlation between city size and the share of Entertainement is positive. We split the sample into pre-1725 and post 1871 to avoid the coexistence of cities from Bairoch sample (where population is at the agglomeration level) and from the Census sample (where population is at the city level). Variables are in log.

Weights used	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$Visibility = transl. \times words$
Mean (sd)	0.32 (0.22)	0.32 (0.22)	0.32 (0.22)	0.28 (0.33)
Largest city	0.24 (0.20)	0.24 (0.20)	0.24 (0.21)	0.18 (0.26)
Second largest city	0.25 (0.24)	0.25 (0.24)	0.25 (0.24)	0.21 (0.30)
Third largest city	0.26 (0.23)	0.26 (0.23)	0.26 (0.23)	0.22 (0.30)
800-1800AD	0.21 (0.21)	0.21 (0.21)	0.21 (0.21)	0.17 (0.27)
1800-1939AD	0.34 (0.18)	0.34 (0.18)	0.34 (0.18)	0.32 (0.33)
1950-2000AD	0.50 (0.18)	0.50 (0.18)	0.50 (0.18)	0.46 (34)

Table 20: Share of population in entertainment in cities, by cells of cities/period (16-63)

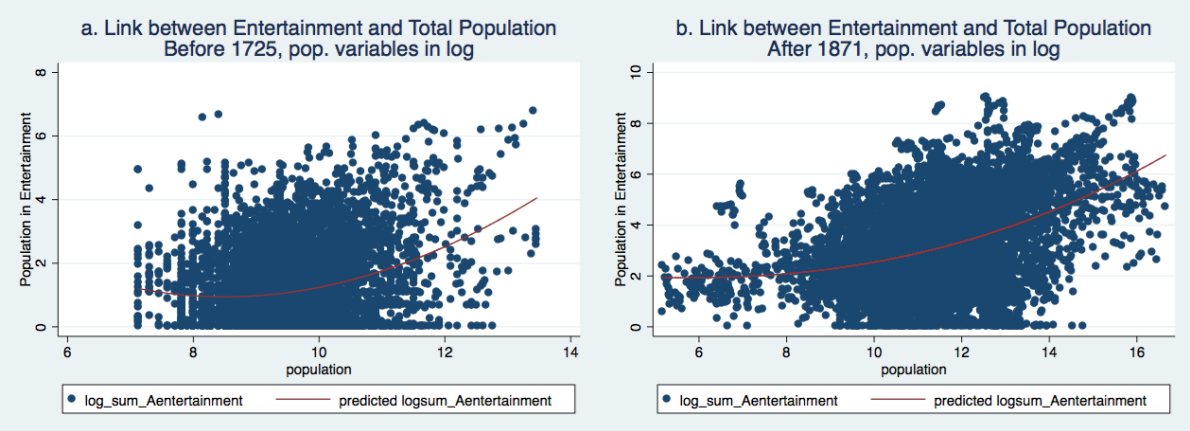


Figure 19: Entertainment and Population

5 Facts on the correlations between city growth and the number of `notable people`

We now proceed to correlation analysis. Our database is organized into time periods, as indicated above: periods of 50 years between 800 and 1700; of 25 years between 1700 and 1800; and 10 years between 1800 and 1939. We exclude the post WWII period from the analysis.

The typical regression is the growth rate of a given variable from one period to the next (population of a city, visibility of `notable people` in that city, number or shares of `notable people` in various occupations in the city) on their own lags (first and second lags), and the lags of other explanatory variables. All regressions are in logs on both sides of the equation and coefficients thus should be interpreted as elasticities. Doing so, we can get a sense of the intrinsic dynamics of the dependent variable (whether there are cycles or regression-to-the-mean) and the lagged cross-correlation between variables. Although it may be tempting to discuss Granger-causality here, we avoid it and only describe coefficients as reflecting the patterns of dynamic cross-correlations.

Table 21 confirms the fact that the visibility of `notable people` overall does not have any positive correlation with population growth, and instead we find here a negative effect, which does not hold

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	City Population Growth Rate					
Log of population $(t-1)$	0.330***	1.305***	0.345***	1.327***	1.304***	1.345***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Log of population $(t-2)$	-0.384***	-0.394***	-0.427***	-0.381***	-0.393***	-0.427***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Log of visibility	-0.017***	-0.010***	-0.003	-	-	-
	(0.003)	(0.003)	(0.003)			
Log of visibility $(t-1)$	-0.004	0.003	0.000	-0.010***	0.002	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log of visibility $(t-2)$	-0.020***	-0.011***	-0.008**	-0.024***	-0.012***	-0.008***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Time Period		-0.008***			-0.006***	
		(0.002)			(0.002)	
Time Period ^{2}		0.000***			0.000***	
		(0.000)			(0.000)	
Constant	0.989***	1.064***	1.034***	0.902***	0.947***	1.005***
	(0.051)	(0.091)	(0.065)	(0.049)	(0.082)	(0.059)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time Period Fixed Effects	No	No	Yes	No	No	Yes
Observations	11,141	11,141	11,141	11,141	11,141	11,141
R-squared	0.521	0.986	0.587	0.986	0.986	0.988

Clustered Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

In the Table, t refers to the aggregate time periods of varying length (50, 25 or 10 years).

Table 21: City growth regressions

once we account for a time period fixed effect in the regression (Equation (3)). Visibility, lagged by two periods is significant at the 5% level but the three variables considered are not siginificant overall. We also find that population lagged by one period has a positive impact on its own growth, but a negative impact when lagged by two periods, indicating a pattern of cycles of length two periods.

Table 22 shows that reciprocally, the visibility of `notable people` is not impacted by the lags of population at least as long as the effect of time periods is neutralized. We also find strong persistence of the visibility variable, which positively depends on its first and second lags.

We now study the effect of the number and shares of `notable people` in each occupation. A key issue here is the respective role of Governance and Entertainment. We restrict our analysis to the period 1800-1939, future versions will study the full period of analysis. Over this time period, we distinguish between two groups of countries, Anglo-Saxon countries (United States, United Kingdom, Australia,

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Mean Log Visibility of `Notable People`					
Log of population	-0.180***	-0.091***	-0.034	-	-	-
	(0.033)	(0.031)	(0.033)			
Log of population $(t-1)$	0.389***	0.139***	0.001	0.151***	0.021	-0.046
	(0.053)	(0.050)	(0.052)	(0.030)	(0.028)	(0.029)
Log of population $(t-2)$	-0.291***	-0.094***	-0.027	-0.222***	-0.059**	-0.013
	(0.030)	(0.028)	(0.029)	(0.027)	(0.025)	(0.025)
Log of visibility $(t-1)$	0.320***	0.147***	0.114***	0.322***	0.147***	0.114***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Log of visibility $(t-2)$	0.189***	0.040***	0.027***	0.194***	0.041***	0.027***
	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)	(0.010)
Constant	5.422***	12.303***	8.603***	5.260***	12.217***	8.569***
	(0.164)	(0.254)	(0.190)	(0.161)	(0.252)	(0.188)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time Period Fixed Effects	No	No	Yes	No	No	Yes
Observations	11,141	11,141	11,141	11,141	11,141	11,141
R-squared	0.699	0.742	0.756	0.698	0.742	0.756

Clustered Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

In the Table, t refers to the aggregate time periods of varying length (50, 25 or 10 years).

Table 22: Visibility Index Regressions

	Anglo-saxon countries	Non anglo-saxon countries
Mean year in the sample	1879	1887
Q25 City population	15 615	23 029
Median City population	48 663	47 071
Q75 City population	127 183	95 186
Share governance	0.438 (0.167)	0.368 (0.193)
Share entertainment	0.310 (0.160)	0.357 (0.190)
Number of cities	252	633
Number of cities / time periods	1,291	3,474

Table 23: Sample statistics over the sample period 1800-1939

New Zealand and Canada) and non-Anglo-Saxon countries. It is noteworthy that at the city level, the samples differ regarding the share of each occupation, with, in particular, a higher share of Governance in Anglo-Saxon countries compensated by a lower share of the Entertainment category. This is not the case on other dimensions such as population and years available, as reported .

We use four specifications where individuals are not weighted, weighted by their percentile of visibility or its square, or finally weighted by their visibility indicator in level. Results are reported in Table 24. Regarding numbers (in logs), not surprisingly, most coefficients are positive, for each of the samples. Interestingly, the (lagged) share of Governance is either negative and significant in the first four columns or positive but marginally significant. Lagged Entertainment is positive, but typically not significant. Finally, and probably the most robust result, the (lagged) share of Entrepreneur is the most positive and significant variable in both sub-samples, with elasticities between 0.03 and 0.01: the first column tells us that a 10% increase in the lagged number of Entrepreneurs is associated (in a non-causal way) with an additional current period population growth of 0.29 percentage points.

A breakdown of the Entertainment category into Sports on the one hand, and Arts/Lit on the other hand, delivers a positive impact of the latter, as a relatively robust result. See Table 25. It remains to be verified and its causal value must not be claimed here.

We finally replicate the analysis using the share of occupations instead of using logs, we select the share of governance as the reference category, since the sum of shares comes to 1 by construction. Further, lagged shares of `notable people` do not provide information on their total number in each city, contrary to the previous table. We therefore further control by their total number, regardless of the category. This number is weighted according to the shares in each specification. Again, Table 26 shows that the lagged share of Entrepreneurs is the most significant variable, especially for Anglo-Saxon countries, while the lagged log number of `notable people` matters for growth in non Anglo-Saxon countries.

Dependent variable	City Population Growth Rate				City Population Growth Rate			
Sample	Anglo-Saxon countries				Non Anglo-Saxon countries			
Weights	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$transl. \times words$	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$transl. \times words$
Log of population $(t-1)$	0.226***	0.230***	0.234***	0.247***	0.285***	0.285***	0.285***	0.289***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.019)	(0.019)	(0.019)	(0.019)
Log of population $(t-2)$	-0.368***	-0.373***	-0.376***	-0.380***	-0.368***	-0.369***	-0.369***	-0.372***
	(0.021)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.020)
Log of entertainment $(t-1)$	-0.001	0.003	0.003	-0.000	0.007	0.007*	0.007**	0.001**
	(0.011)	(0.010)	(0.008)	(0.001)	(0.004)	(0.004)	(0.003)	(0.001)
Log of governance $(t-1)$	-0.067***	-0.051***	-0.036***	-0.003**	0.008**	0.006*	0.006*	0.002***
	(0.015)	(0.013)	(0.010)	(0.001)	(0.004)	(0.004)	(0.003)	(0.001)
Log of academics $(t-1)$	0.006	0.005	0.006	0.002	0.008*	0.007*	0.006*	0.000
	(0.011)	(0.009)	(0.008)	(0.001)	(0.004)	(0.004)	(0.003)	(0.001)
Log of entrepreneur $(t-1)$	0.029**	0.023**	0.017**	0.000	0.009*	0.010**	0.009***	0.002***
	(0.011)	(0.010)	(0.008)	(0.001)	(0.005)	(0.004)	(0.003)	(0.001)
Log of family $(t-1)$	0.015**	0.010	0.006	0.001	0.001	0.001	0.000	0.000
	(0.008)	(0.007)	(0.005)	(0.001)	(0.005)	(0.004)	(0.003)	(0.000)
Log of other $(t-1)$	0.017**	0.008	0.005	0.000	-0.017***	-0.012***	-0.009***	-0.001**
	(0.008)	(0.007)	(0.006)	(0.001)	(0.005)	(0.004)	(0.003)	(0.000)
Constant	1.866***	1.827***	1.770***	1.630***	1.015***	1.018***	1.021***	1.009***
	(0.129)	(0.128)	(0.125)	(0.118)	(0.100)	(0.100)	(0.100)	(0.100)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Period Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,291	1,291	1,291	1,291	3,474	3,474	3,474	3,474
R-squared	0.787	0.785	0.784	0.782	0.580	0.580	0.581	0.581

Clustered Standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1.

In the Table, t refers to the aggregate time periods of varying length (50, 25 or 10 years).

Table 24: City growth regressions based on the number of `notable people` per category

Dependent variable	City Population Growth Rate				City Population Growth Rate			
Sample	Anglo-Saxon countries				Non Anglo-Saxon countries			
Weights	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$transl. \times words$	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$transl. \times words$
Log of population $(t-1)$	0.223***	0.227***	0.233***	0.249***	0.273***	0.273***	0.273***	0.281***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.019)	(0.019)	(0.019)	(0.019)
Log of population $(t-2)$	-0.372***	-0.376***	-0.378***	-0.380***	-0.359***	-0.359***	-0.360***	-0.365***
	(0.021)	(0.021)	(0.021)	(0.021)	(0.020)	(0.019)	(0.019)	(0.020)
Log of arts/lit/media $(t-1)$	0.026**	0.026***	0.020**	0.001	0.012***	0.011***	0.009***	0.001**
	(0.011)	(0.009)	(0.008)	(0.001)	(0.004)	(0.004)	(0.003)	(0.001)
Log of sports $(t-1)$	-0.007	-0.009	-0.009	-0.006	-0.025***	-0.025***	-0.024***	-0.021***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.003)
Log of governance $(t-1)$	-0.071***	-0.055***	-0.040***	-0.003**	0.007*	0.005	0.005	0.002**
	(0.015)	(0.013)	(0.010)	(0.001)	(0.004)	(0.004)	(0.003)	(0.001)
Log of academics $(t-1)$	-0.001	-0.000	0.003	0.002	0.007*	0.007*	0.006*	0.001
	(0.011)	(0.010)	(0.008)	(0.001)	(0.004)	(0.004)	(0.003)	(0.001)
Log of entrepreneur $(t-1)$	0.023**	0.020**	0.016**	0.000	0.013***	0.014***	0.013***	0.002***
	(0.011)	(0.010)	(0.008)	(0.001)	(0.005)	(0.004)	(0.003)	(0.001)
Log of family $(t-1)$	0.014*	0.009	0.006	0.001	0.001	0.001	0.001	0.000
	(0.008)	(0.007)	(0.005)	(0.001)	(0.005)	(0.004)	(0.003)	(0.000)
Log of other $(t-1)$	0.016**	0.008	0.005	0.000	-0.010**	-0.006	-0.004	-0.001*
	(0.008)	(0.007)	(0.006)	(0.001)	(0.005)	(0.004)	(0.003)	(0.000)
Constant	1.934***	1.884***	1.808***	1.613***	1.062***	1.065***	1.065***	1.037***
	(0.132)	(0.130)	(0.126)	(0.119)	(0.100)	(0.100)	(0.100)	(0.100)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Period Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,291	1,291	1,291	1,291	3,474	3,474	3,474	3,474
R-squared	0.788	0.787	0.786	0.782	0.587	0.588	0.588	0.586

Clustered Standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1.

In the Table, t refers to the aggregate time periods of varying length (50, 25 or 10 years).

Table 25: City growth regressions based on the number of `notable people` per category

Dependent variable	City Population Growth Rate				City Population Growth Rate			
Sample	Anglo-Saxon countries				Non Anglo-Saxon countries			
Weights	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$transl. \times words$	Unweighted	$(2 \times pctile/100)$	$(2 \times pctile/100)^2$	$transl. \times words$
Log of population $(t-1)$	0.228***	0.231***	0.236***	0.249***	0.286***	0.286***	0.287***	0.290***
	(0.024)	(0.025)	(0.025)	(0.024)	(0.019)	(0.019)	(0.019)	(0.019)
Log of population $(t-2)$	-0.367***	-0.373***	-0.376***	-0.379***	-0.371***	-0.371***	-0.371***	-0.372***
	(0.021)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)	(0.020)	(0.020)
Share entertainment $(t-1)$	0.062	0.076	0.070	0.004	0.007	0.009	0.008	-0.003
	(0.064)	(0.057)	(0.048)	(0.012)	(0.013)	(0.013)	(0.013)	(0.006)
Share academics $(t-1)$	0.223**	0.173*	0.149**	0.021	0.017	0.017	0.016	0.001
	(0.111)	(0.091)	(0.073)	(0.016)	(0.017)	(0.017)	(0.016)	(0.009)
Share entrepreneur $(t-1)$	0.361***	0.297***	0.227***	-0.004	0.028	0.040	0.046*	0.032*
	(0.092)	(0.081)	(0.070)	(0.019)	(0.028)	(0.028)	(0.027)	(0.017)
Share family $(t-1)$	0.434	0.404	0.310	0.006	0.011	0.032	0.041	0.018
	(0.348)	(0.290)	(0.235)	(0.046)	(0.058)	(0.055)	(0.052)	(0.021)
Share other $(t-1)$	0.745***	0.367**	0.212	0.045	-0.109**	-0.080*	-0.063	-0.066***
	(0.217)	(0.186)	(0.153)	(0.037)	(0.049)	(0.047)	(0.044)	(0.023)
Share governance (ref. category)	-	-	-	-	-	-	-	-
Log number of `notable people` $(t-1)$	-0.013	-0.011	-0.010	-0.003	0.010***	0.010***	0.009***	0.003***
	(0.010)	(0.010)	(0.009)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Constant	1.675***	1.692***	1.685***	1.616***	1.055***	1.047***	1.041***	0.994***
	(0.117)	(0.119)	(0.119)	(0.118)	(0.101)	(0.101)	(0.101)	(0.100)
City Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Period Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,291	1,291	1,291	1,291	3,474	3,474	3,474	3,474
R-squared	0.788	0.786	0.784	0.781	0.580	0.579	0.579	0.580

Clustered Standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1.

In the Table, t refers to the aggregate time periods of varying length (50, 25 or 10 years).

Table 26: City growth regressions based on the shares of `notable people` per category

6 Conclusion

This paper is a first step into using long-run historical data from the Internet to conduct an economic analysis of the city growth. Future iterations will eliminate remaining errors in data and extend the empirical analysis beyond correlations to provide an assessement of causal links.

References

- de la Croix David and Licandro Omar. (2015). The longevity of famous people from Hammurabi to Einstein. *Journal of Economic Growth*, volume 20(3), doi 10.1007/s10887-015-9117-0, pages 263-303.

- Schich Maximilian, Song Chaoming, Ahn Yong-Yeol, Mirsky Alexander, Martino Mauro, Barabási Albert-László and Dirk Helbing. (2014). A network framework of cultural history. *Science*, 345, 6196, pages 558-562

- Yu Amy Zhao, Ronen Shahar, Hu Kevin, Lu Tiffany and César A. Hidalgo. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data,* 3, Article number: 150075.

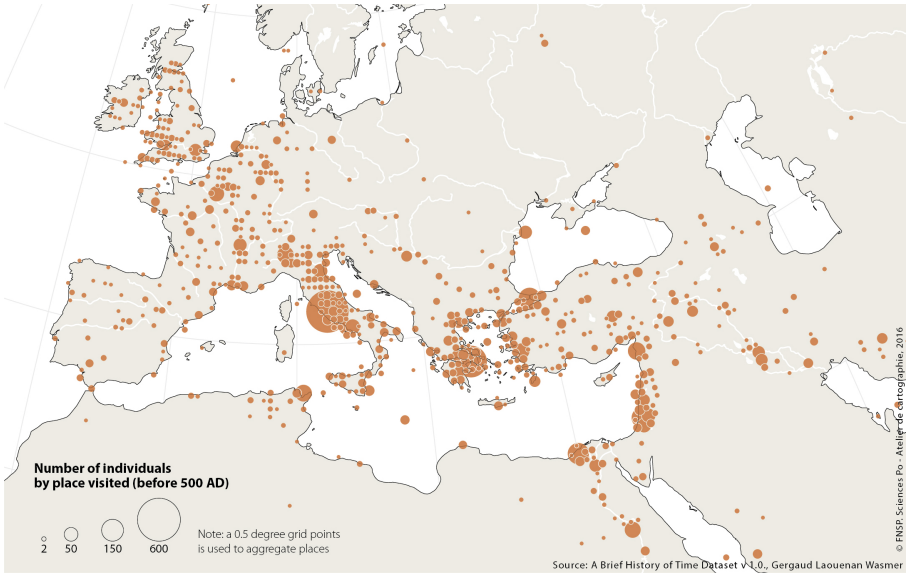


Figure A.1: Number of individuals by places (before 500)

A Appendix

A.1 MAPS

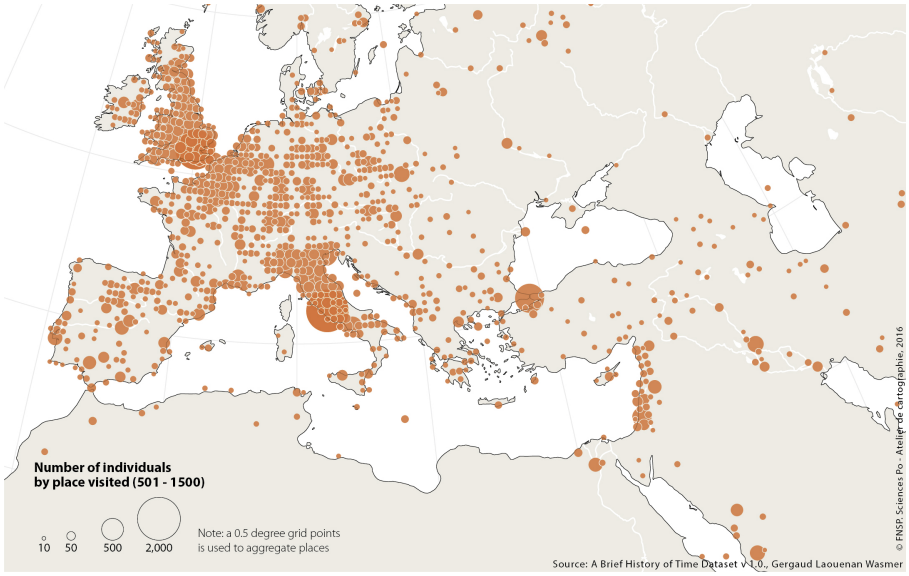


Figure A.2: Number of individuals by places (501-1500)

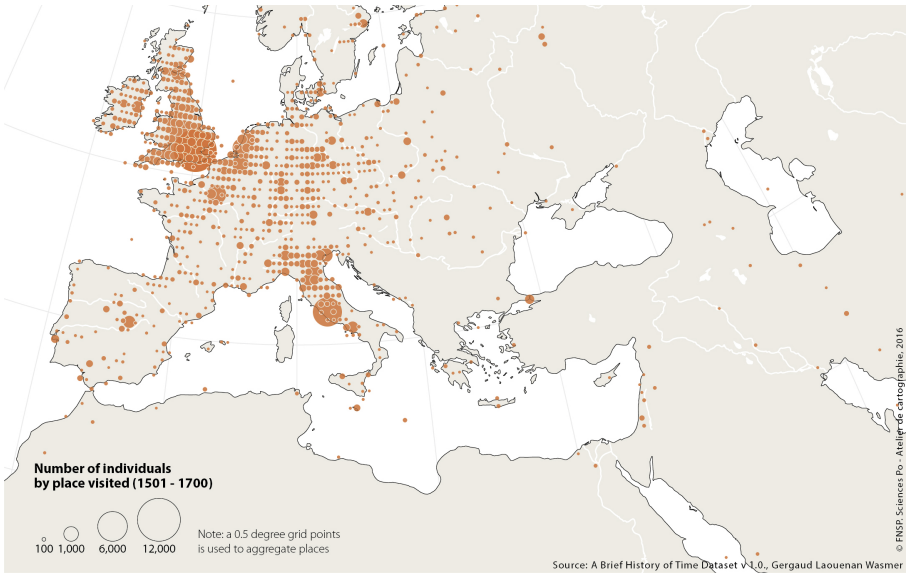


Figure A.3: Number of individuals by places (1501-1700)

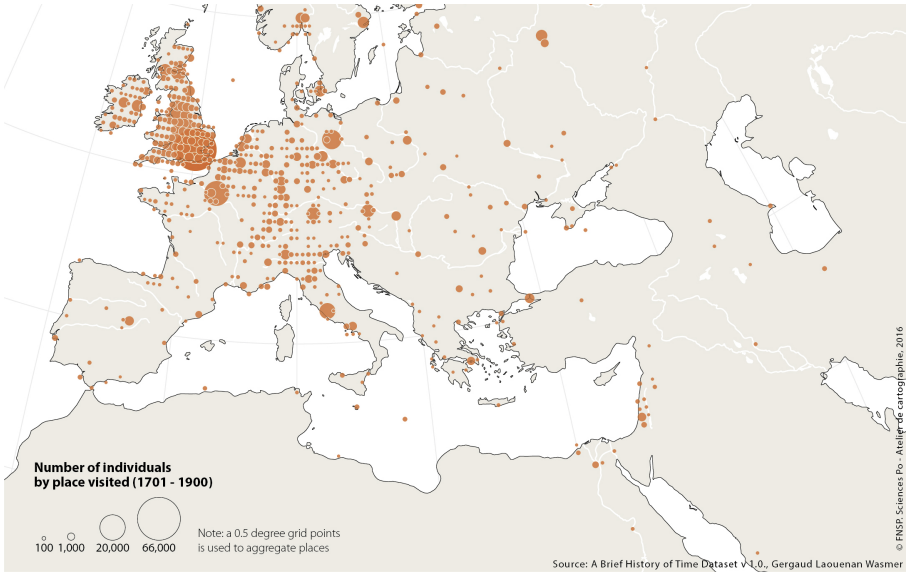


Figure A.4: Number of individuals by places (1701-1900)

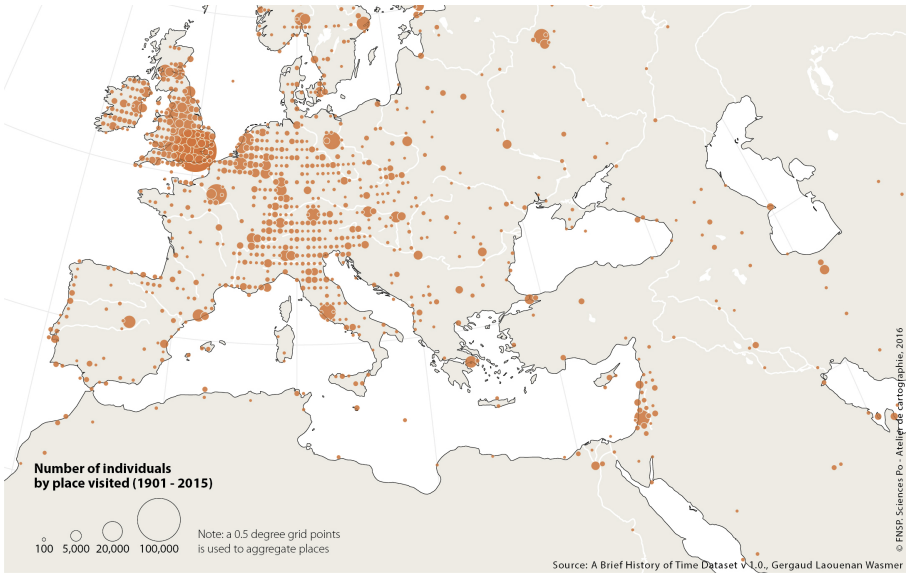


Figure A.5: Number of individuals by places (1901-2015)

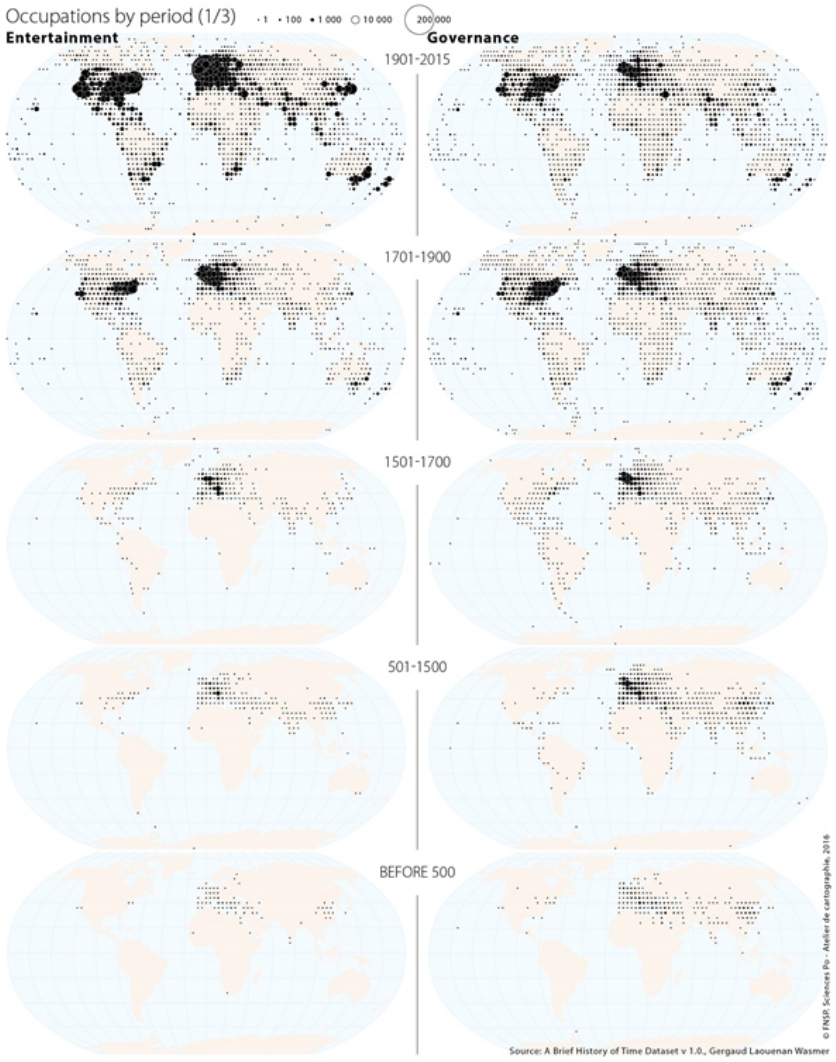


Figure A.6: Entertainment & Governance

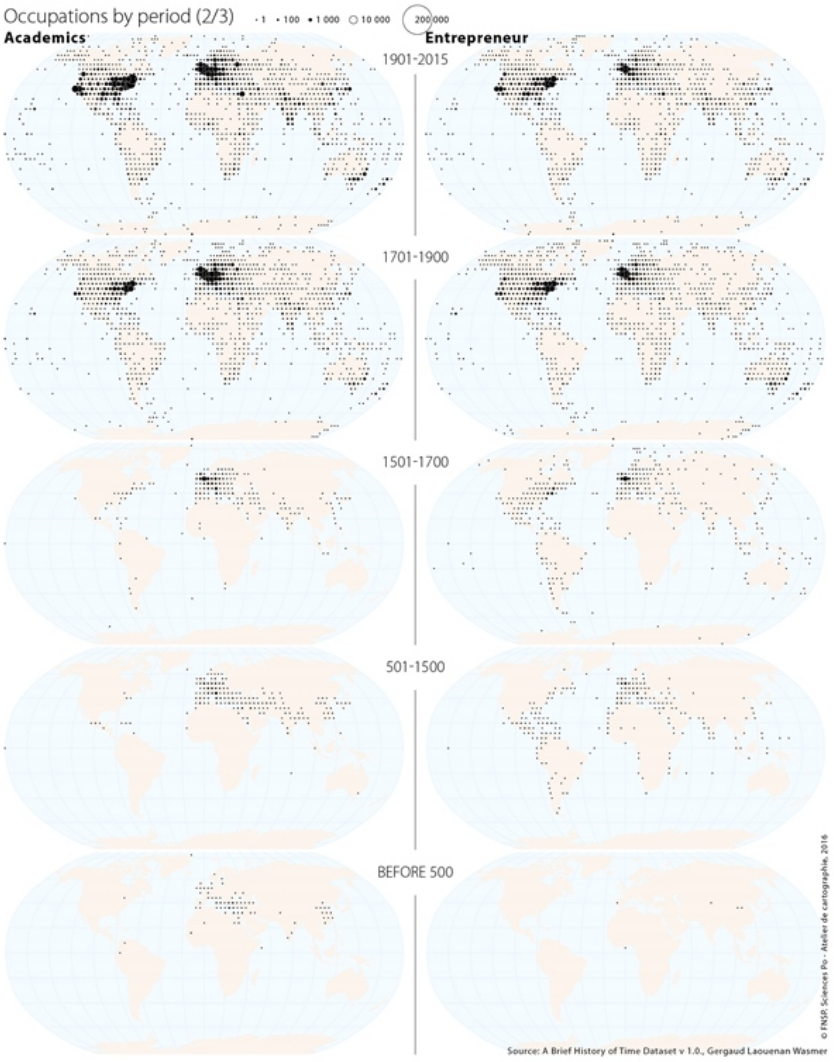


Figure A.7: Academics & Entrepreneur

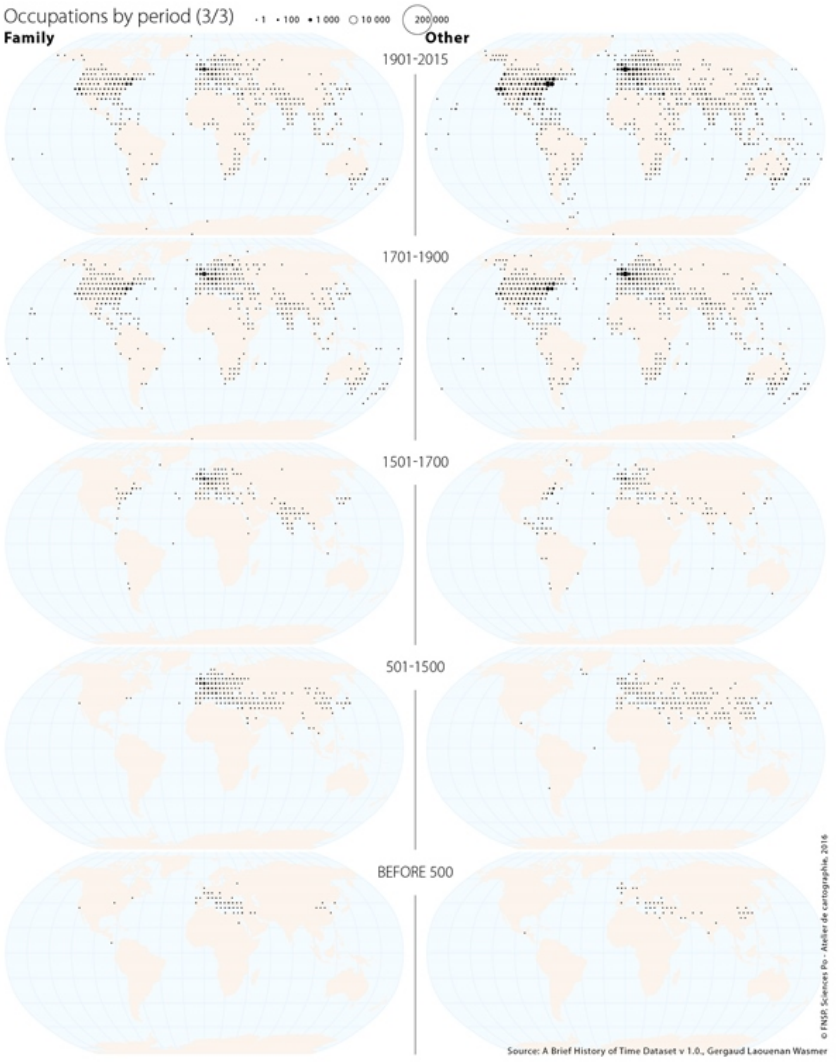


Table A.1: Family & Other

A.2 People by Occupations

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	32	Karl Marx	1818	1883	Germany	studies	studies
2	189	David Hume	1711	1776	Scotland	studies	studies
3	193	John Maynard Keynes	1883	1946	England	studies	Other
4	423	Milton Friedman	1912	2006	US	studies	studies
5	667	Friedrich Hayek	1899	1992	Austria	studies	studies
6	698	Muhammad Yunus	1940		Bangladesh	business	business
7	894	Paul Krugman	1953		US	studies	education
8	1024	Joseph Stiglitz	1943		US	studies	education
9	1668	Amartya Sen	1933		India	studies	studies
10	2062	Peter Kropotkin	1842	1921	Russia	studies	studies
11	2200	Paul Samuelson	1915	2009	US	studies	studies
12	2279	Chanakya	-370	-283	India	education	studies
13	2688	Gary Becker	1930	2014	US	studies	education
14	2761	Murray Rothbard	1926	1995	US	studies	studies
15	2876	Kenneth Arrow	1921		US	studies	lit
16	2913	Leonid Hurwicz	1917	2008	Poland	studies	studies
17	3365	Alan Greenspan	1926		US	studies	business
18	3825	Ben Bernanke	1953		US	studies	business
19	4242	Shaukat Aziz	1949		Pakistan	studies	lit
20	4265	Ronald Coase	1910	2013	England	studies	lit
			Top decile (random sample)				
269	98711	Jean Fourastié	1907	1990	France	studies	lit
270	98722	Franz Hermann Schulze-Delitzsch	1808	1883	Germany	studies	studies
271	99463	Iuliu Winkler	1964		Romania	business	studies
272	99511	Ladislaus Bortkiewicz	1868	1931	Russia	studies	studies
273	99995	Steve Keen	1953		Australia	studies	lit
			First quartile (random sample)				
676	292055	Alvin Saunders Johnson	1874	1971	US	studies	Other
677	292200	Peyton Young	1945		US	sports	studies
678	292649	Cecilia Rouse	1963		US	studies	education
680	293982	Takashi Negishi	1933		Japan	studies	studies
679	294001	Georg Friedrich Sartorius	1765	1828	Germany	studies	studies
			Median (random sample)				
1354	566073	Svetlana Kirdina	1955		Russia	studies	studies
1353	566131	Gabibulla Rabadanovich Khasaev	1951		Russia	studies	education
1355	566691	Maria Kiwanuka	1955		Uganda	studies	business
1356	566822	Jan Kregel	1944		US	studies	studies
1357	567479	Claudio Dematté	1942	2004	Italy	studies	Other
			Third quartile (random sample)				
2030	855631	Dattatreya Gopal Karve	1898	1967	India	studies	education
2031	855881	Thomas Mayer (American economist)	1927	2015	US	studies	education
2032	856632	Alfredo Salazar			.	studies	education
2033	856888	Ray Major	1961		.	studies	studies
2034	856947	Wim Driehuis	1943		Netherlands	studies	education
			Last decile (random sample)				
2437	1046320	Gaspar Roca	1926	2007	US	lit	studies
2438	1048363	Henryk Kierzkowski	1943		Poland	studies	studies
2439	1048843	Ken-Ichi Inada	1925	2002	Japan	studies	studies
2441	1048902	Edward Wolff	1946		US	studies	education
2440	1048976	Reetika Khera			India	studies	studies

Table A.2: Economists (excluding the governance category)

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	16	Cristiano Ronaldo	1985		Portugal	sports	sports
2	22	Roger Federer	1981		Switzerland	sports	sports
3	24	Lionel Messi	1987		Argentina	sports	sports
4	25	Novak Djokovic	1987		Serbia	sports	sports
5	38	Rafael Nadal	1986		Spain	sports	sports
6	72	Serena Williams	1981		US	sports	sports
7	93	Bobby Fischer	1943	2008	US	sports	Other
8	97	Michael Schumacher	1969		Germany	sports	sports
9	99	David Beckham	1975		England	sports	sports
10	102	Pelé	1940		Brazil	sports	sports
11	106	Michael Phelps	1985		US	sports	sports
12	108	Michael Jordan	1963		US	sports	sports
13	123	Maria Sharapova	1987		Russia	sports	sports
14	127	Diego Maradona	1960		Argentina	sports	sports
15	136	Francesco Totti	1976		Italy	sports	sports
16	165	Andy Murray	1987		Scotland	sports	sports
17	180	Fernando Alonso	1981		Spain	sports	sports
18	183	Muhammad Ali	1942		US	sports	sports
19	186	Usain Bolt	1986		Jamaica	sports	sports
20	188	Kobe Bryant	1978		US	sports	sports
			Top decile (random sample)				
40816	125315	Rudolf Bester	1983		Namibia	sports	sports
40815	125320	Gimax	1938		Italy	sports	sports
40812	125321	Mario Ghella	1929		Italy	sports	sports
40814	125329	Nikolaj Ehlers	1996		Denmark	sports	sports
40813	125331	Marcell Deák-Nagy	1992		Hungary	sports	sports
			First quartile (random sample)				
102033	300267	Ben Eaves	1982		US	sports	sports
102035	300269	Harald Smith	1879	1977	Norway	sports	sports
102034	300284	Ginger Molloy	1937		New Zealand	sports	sports
102036	300290	Paul Gruber	1965		US	sports	sports
			Median (random sample)				
204067	611202	Iosif Anisim			Romania	sports	sports
204070	611203	Ivette María	1975		Spain	sports	sports
204069	611214	Arantxa Sanchis	1990		India	sports	sports
204068	611266	Hamid Reza Fathi	1980		Iran	sports	sports
			Third quartile (random sample)				
306104	944207	Karl Esleeck	1903	1952	US	education	sports
306102	944237	Orian Landreth	1904	1996	US	sports	sports
306101	944369	Bill Richardson (footballer born 1943)	1943		England	sports	sports
306103	944402	Joe Blythe	1881		England	sports	sports
			Last decile (random sample)				
367323	1125362	George Barron	1883	1961	England	sports	sports
367324	1125650	Don Lisbon	1941		Canada	sports	sports
367322	1125676	Aleksandr Dementyev	1995		Russia	sports	sports
367321	1125800	Gordon Mair	1958		Scotland	sports	sports
367325	1125810	Vlad Negoitescu	1991		Romania	sports	sports

Table A.3: Individuals in Sports category

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	26	Charlie Chaplin	1889	1977	England	arts	arts
2	43	Vincent Van Gogh	1853	1890	Netherlands	arts	arts
3	44	Johann Sebastian Bach	1685	1750	Germany	arts	arts
4	57	Leonardo da Vinci	1452	1519	Italy	inventor	arts
5	114	Rabindranath Tagore	1861	1941	India	lit	arts
6	118	Ludwig Van Beethoven	1770	1827	Germany	arts	arts
7	139	Richard Wagner	1813	1883	Germany	arts	arts
8	141	Mark Twain	1835	1910	US	lit	arts
9	144	Pablo Picasso	1881	1973	Spain	arts	arts
10	152	Frédéric Chopin	1810	1849	Poland	arts	arts
11	158	Wolfgang Amadeus Mozart	1756	1791	.	arts	Family
12	159	Michelangelo	1475	1564	Italy	arts	arts
13	163	Frank Lloyd Wright	1867	1959	US	arts	arts
14	217	Giuseppe Verdi	1813	1901	Italy	arts	arts
15	250	Antoni Gaudí	1852	1926	Spain	arts	arts
16	260	Paul Gauguin	1848	1903	France	arts	arts
17	268	Pyotr Ilyich Tchaikovsky	1840	1893	Russia	arts	arts
18	287	Rembrandt	1606	1669	Netherlands	arts	arts
19	391	Raphael	1483	1520	Italy	arts	arts
20	421	Gustav Mahler	1860	1911	Austria	arts	arts
			Top decile (random sample)				
4397	109405	Henry Lytton	1865	1936	England	arts	arts
4398	109411	Adolphe d'Ennery	1811	1899	France	arts	lit
4400	109450	Marguerite Durand	1864	1936	France	arts	lit
4399	109456	Gioseffo Guami	1542	1611	Italy	arts	arts
			First quartile (random sample)				
10996	250344	Ádám Récsey	1775	1852	Hungary	arts	politics
10994	250366	William Thoms	1803	1885	England	lit	arts
10995	250385	Twm o'r Nant	1739	1810	Wales	arts	lit
10997	250434	Henri Gagnebin	1886	1977	Belgium	arts	arts
			Median (random sample)				
21989	502347	Peter Ferdinand Funck	1788	1859	Denmark	arts	arts
21990	502381	William Adams Delano	1874	1960	US	arts	business
21991	502591	Adam Eberle	1804	1832	.	arts	arts
21992	502592	Louis Round Wilson	1876	1979	US	studies	arts
			Third quartile (random sample)				
32985	835167	Sydney Mitchell	1856	1930	Scotland	arts	arts
32984	835236	Umberto Coromaldi	1870	1948	Italy	arts	arts
32987	835288	Olive Mudie-Cooke	1890	1925	England	arts	arts
32986	835328	Lucy Isabella Buckstone	1859	1893	England	arts	arts
			Last decile (random sample)				
39584	1096396	Giovan Giacomo Dalla Corna	1480	1560	Italy	arts	business
39582	1096495	Arthur Hampson	1878	1952	England	sports	arts
39581	1096614	James Leslie Findlay	1868	1952	Scotland	arts	military
39583	1096908	Mary Elizabeth Turner Salter	1856	1938	US	arts	arts

Table A.4: Arts/Litt born before 1890

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	8	Michael Jackson	1958	2009	US	arts	arts
2	35	Elvis Presley	1935	1977	US	arts	arts
3	56	Paul McCartney	1942		England	arts	arts
4	59	Lady Gaga	1986		US	arts	arts
5	66	Bob Dylan	1941		US	arts	arts
6	68	Frank Sinatra	1915	1998	US	arts	arts
7	74	Akira Kurosawa	1910	1998	Japan	arts	arts
8	78	Celine Dion	1968		Canada	arts	arts
9	80	Madonna (entertainer)	1958		US	arts	arts
10	87	Marilyn Monroe	1926	1962	US	arts	arts
11	94	Whitney Houston	1963	2012	US	arts	arts
12	96	Beyoncé Knowles	1981		US	arts	arts
13	98	Taylor Swift	1989		US	arts	arts
14	107	Eminem	1972		US	arts	arts
15	109	John Lennon	1940	1980	England	arts	arts
16	113	Meryl Streep	1949		US	arts	arts
17	131	Stanley Kubrick	1928	1999	US	arts	lit
18	132	Angelina Jolie	1975		US	arts	arts
19	135	Rihanna	1988		Barbados	arts	arts
20	145	Steven Spielberg	1946		US	arts	lit
			Top decile (random sample)				
19149	83815	Uhm Ji-Won	1977		South Korea	arts	arts
19150	83822	Satoshi Urushihara	1966		Japan	arts	arts
19152	83830	Jeremy Jordan (stage actor)	1984		US	arts	arts
19151	83832	Yasumi Matsuno	1965		Japan	sports	arts
19153	83849	Eileen Joyce	1908	1991	Australia	arts	arts
			First quartile (random sample)				
47876	232537	Oleg Strizhenov	1929		Russia	arts	arts
47875	232545	Meg Mundy	1915		England	arts	arts
47874	232561	Go Eun-Mi	1976		South Korea	arts	arts
47878	232567	Jamal Rahimov	1987		Azerbaijan	sports	arts
47877	232575	Andrea Rost	1962		Hungary	arts	arts
			Median (random sample)				
95750	495326	Tully Satre	1989		US	arts	arts
95752	495333	Claire Falkenstein	1908	1997	US	arts	arts
95751	495339	Rick Krebs	1949		.	sports	arts
95753	495405	Henry Johnson (guitarist)	1954		US	arts	arts
			Third quartile (random sample)				
143629	804086	Brett Goldsmith	1961		Australia	arts	politics
143628	804119	Peter Gvozdják	1965		Slovakia	sports	arts
143627	804147	Carl Von Hanno	1901	1953	Norway	arts	politics
143625	804188	John Wylie (musician)	1974		US	arts	arts
143626	804303	George Dahl	1894	1987	US	arts	arts
			Last decile (random sample)				
172352	1027662	Jamie Redfern	1957		England	lit	arts
172353	1027847	Lenin M. Sivam	1974		Canada	arts	arts
172354	1027862	Paul Meehan	1938		England	sports	arts
172351	1027987	Chad Connell	1983		Canada	arts	arts
172350	1028035	Anindita Nayar	1988		India	arts	arts

Table A.5: Arts/Litt born after 1890

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	1	Barack Obama	1961		US	politics	education
2	2	Ronald Reagan	1911	2004	US	politics	arts
3	4	George W. Bush	1946		US	politics	business
4	5	Napoleon Bonaparte	1769	1821	France	military	politics
5	6	Winston Churchill	1874	1965	England	politics	politics
6	7	Nelson Mandela	1918	2013	South Africa	politics	politics
7	9	Adolf Hitler	1889	1945	Austria	politics	politics
8	10	Joseph Stalin	1878	1953	Russia	politics	politics
9	11	John F. Kennedy	1917	1963	US	politics	politics
10	12	Mahatma Gandhi	1869	1948	India	politics	politics
11	17	Pope John Paul Ii	1920	2005	Poland	politics	religious
12	20	Franklin D. Roosevelt	1882	1945	US	politics	politics
13	21	Abraham Lincoln	1809	1865	US	politics	politics
14	23	George Washington	1732	1799	US	politics	military
15	27	Vladimir Lenin	1870	1924	Russia	politics	politics
16	28	Charles de Gaulle	1890	1970	France	lit	politics
17	29	Hillary Rodham Clinton	1947		US	politics	politics
18	31	Bill Clinton	1946		US	politics	politics
19	33	Pope Benedict Xvi	1927		Germany	politics	religious
20	34	Che Guevara	1928	1967	Argentina	politics	politics
		Top decile (random sample)					
22359	155389	M. Kulasegaran	1957		Malaysia	politics	politics
22360	155414	Daniel Pfeiffer	1975		US	politics	politics
22361	155423	Gheorghe Cristescu	1882	1973	Romania	politics	politics
22362	155436	Augusto del Noce	1910	1989	Italy	studies	politics
		First quartile (random sample)					
55900	378806	Anthony Enahoro	1923	2010	Nigeria	politics	politics
55901	378850	Jorge del Prado Chávez	1910	1999	Peru	politics	politics
55902	378871	Vera Chirwa	1932		Malawi	law	politics
55899	378916	Salvador María del Carril	1798	1883	Argentina	law	politics
		Median (random sample)					
111799	691320	Robert S. Hall	1879	1941	US	politics	politics
111800	691396	Jørgen Flood	1792	1867	Norway	business	politics
111801	691464	Jessie Gruman	1953	2014	.	Other	politics
111802	691483	Zhu Shaolian	1887	1929	China	business	politics
		Third quartile (random sample)					
167699	992734	Richard Bright (politician)	1822	1878	England	politics	politics
167700	992832	Roswell G. Ham	1891	1983	US	education	politics
167702	992906	Jeffrey Herbst	1961		US	politics	studies
167701	993044	Ashley Goldsworthy			Australia	politics	politics
		Last decile (random sample)					
201240	1161643	Solomon Quetsch	1798	1856	Austria	religious	politics
201241	1161787	François-Xavier Méthot	1796	1853	Canada	business	politics
201242	1161914	Mukunda Ram Choudhury			India	politics	law
201239	1161972	William Nash (Manitoba politician)	1846	1917	Canada	law	politics

Table A.6: Governance

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	14	William Shakespeare	1564	1616	England	lit	lit
2	70	Friedrich Nietzsche	1844	1900	Germany	studies	lit
3	75	Jean-Jacques Rousseau	1712	1778	.	studies	lit
4	81	Franz Kafka	1883	1924	Austria	lit	lit
5	90	Voltaire	1694	1778	France	lit	studies
6	114	Rabindranath Tagore	1861	1941	India	lit	arts
7	120	Charles Dickens	1812	1870	England	lit	lit
8	121	Confucius	-551	-479	China	education	lit
9	141	Mark Twain	1835	1910	US	lit	arts
10	142	Fyodor Dostoyevsky	1821	1881	Russia	lit	lit
11	170	Avicenna	980	1037	.	lit	studies
12	225	Oscar Wilde	1854	1900	Ireland	lit	lit
13	239	Victor Hugo	1802	1885	France	lit	lit
14	263	Leo Tolstoy	1828	1910	Russia	lit	lit
15	267	George Bernard Shaw	1856	1950	Ireland	lit	lit
16	281	James Joyce	1882	1941	Ireland	lit	lit
17	288	Rudyard Kipling	1865	1936	England	lit	lit
18	303	Miguel de Cervantes	1547	1616	Spain	lit	lit
19	304	Baruch Spinoza	1632	1677	Netherlands	studies	lit
20	308	Jules Verne	1828	1905	France	lit	lit
			Top decile (random sample)				
2056	83546	Mikha'il Na'ima	1889	1988	Lebanon	lit	lit
2055	83561	Christopher Anstey	1724	1805	England	lit	lit
2057	83604	Joseph Glanvill	1636	1680	England	lit	studies
2058	83722	Paul Fort	1872	1960	France	lit	studies
2059	83781	Jacques Bainville	1879	1936	France	studies	lit
			First quartile (random sample)				
5139	225273	Minamoto no Shunrai	1055	1129	Japan	lit	Family
5141	225296	Dulduityn Danzanravjaa	1803	1856	Mongolia	lit	arts
5140	225312	Jean-Baptiste-Antoine Suard	1733	1817	France	lit	studies
5142	225331	Sara Jeannette Duncan	1861	1922	Canada	lit	lit
5143	225344	Thomas Muir (mathematician)	1844	1934	Scotland	studies	lit
			Median (random sample)				
10280	496058	John Paget (author)	1808	1892	England	worker	lit
10279	496073	Tudur Aled	1460	1525	Wales	lit	lit
10281	496492	Edmond Tarbé Des Sablons	1838	1900	France	lit	lit
10282	496516	Rosauro Almario	1886	1933	Philippines	lit	lit
10283	496618	Étienne Weill-Raynal	1887	1982	France	studies	lit
			Third quartile (random sample)				
15422	846629	Edward Anthony (photographer)	1819	1888	US	lit	Other
15419	846660	Kathleen Hawkins	1883	1981	New Zealand	lit	business
15420	846688	Elizabeth Wynne Fremantle	1778	1857	.	lit	Family
15421	846723	John E. Tullidge	1806	1873	US	arts	lit
15423	846970	Frank Sayers	1763	1817	England	lit	lit
			Last decile (random sample)				
18505	1062973	Jean Middlemass	1833	1919	England	lit	lit
18507	1063028	Samuel Johnson Jr.	1757	1836	England	lit	education
18506	1063130	Lydia Mackenzie Falconer Miller	1812	1876	England	lit	familyB
18504	1063168	William C. McClintock	1845		US	lit	lit
18503	1063290	John Shirreff	1759	1818	Scotland	lit	lit

Table A.7: Litterature born before 1891

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	30	Albert Einstein	1879	1955	Germany	studies	studies
2	32	Karl Marx	1818	1883	Germany	studies	studies
3	45	Sigmund Freud	1856	1939	Austria	studies	studies
4	49	Isaac Newton	1642	1727	England	studies	studies
5	54	Immanuel Kant	1724	1804	Germany	studies	studies
6	58	Charles Darwin	1809	1882	England	studies	studies
7	63	Galileo Galilei	1564	1642	Italy	studies	studies
8	70	Friedrich Nietzsche	1844	1900	Germany	studies	lit
9	75	Jean-Jacques Rousseau	1712	1778	.	studies	lit
10	77	Søren Kierkegaard	1813	1855	Denmark	studies	religious
11	90	Voltaire	1694	1778	France	lit	studies
12	101	Bertrand Russell	1872	1970	England	studies	studies
13	130	Gottfried Wilhelm Von Leibniz	1646	1716	Germany	studies	studies
14	133	Carl Linnaeus	1707	1778	Sweden	studies	studies
15	140	Alan Turing	1912	1954	England	business	studies
16	146	Neil Armstrong	1930	2012	US	studies	studies
17	147	Socrates	-470	-390	Greece	studies	Other
18	160	Alexander Graham Bell	1847	1922	England	studies	inventor
19	170	Avicenna	980	1037	.	lit	studies
20	174	Marie Curie	1867	1934	Poland	studies	studies
			Top decile (random sample)				
9211	112329	Liu E	1857	1909	China	lit	studies
9212	112353	Stanley Coren	1942		.	education	studies
9213	112378	Franciscus Junius (the younger)	1591	1677	Germany	business	studies
9214	112395	Thomas Thomson (chemist)	1773	1852	Scotland	studies	studies
9215	112401	Boris Rybakov	1908	2001	Russia	studies	studies
			First quartile (random sample)				
23030	278169	Susanne Nyström	1982		Sweden	sports	studies
23031	278270	Peter Nicholson (architect)	1765	1844	England	arts	studies
23032	278304	Albert Pilát	1903	1974	Czech Republic	studies	studies
23034	278321	David G. Hartwell	1941		US	lit	studies
23033	278329	Alfred Twardecki	1962		Poland	studies	studies
			Median (random sample)				
46063	550855	George Molnar (philosopher)	1934	1991	Hungary	studies	education
46064	550857	Josselyn Van Tyne	1902	1957	US	studies	arts
46062	550877	Gheorghe Pintilie	1902	1985	Russia	studies	politics
46065	550900	Clayton Oscar Person	1922	1990	Canada	lit	studies
			Third quartile (random sample)				
69094	874446	Henry Jay Forman			US	education	studies
69095	874460	Vandana Singh			India	studies	lit
69097	874497	John Tosh			England	studies	education
69093	874619	John Roberton (1776)	1776	1840	Scotland	studies	politics
69096	874638	Philip D. Morgan	1949		England	studies	inventor
			Last decile (random sample)				
82912	1079380	Rod Coutts			Canada	education	studies
82913	1079474	Robert Bruegmann			.	studies	arts
82916	1079528	Oreste Piro	1954		Argentine	studies	studies
82915	1079630	Molly Worthen	1981		US	studies	lit
82914	1079633	Taylor Carman			US	studies	education

Table A.8: Science

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	5	Napoleon Bonaparte	1769	1821	France	military	politics
2	19	Mustafa Kemal Atatürk	1881	1938	Turkey	military	military
3	23	George Washington	1732	1799	US	politics	military
4	42	Dwight D. Eisenhower	1890	1969	US	politics	military
5	79	Ulysses S. Grant	1822	1885	US	politics	military
6	104	Osama Bin Laden	1957	2011	Saudi Arabia	Other	military
7	210	Chiang Kai-Shek	1887	1975	China	politics	military
8	218	Bashar Al-Assad	1965		Syria	politics	military
9	219	Erwin Rommel	1891	1944	Germany	military	military
10	237	Francisco Franco	1892	1975	Spain	politics	military
11	295	Arthur Wellesley, 1st Duke of Wellington	1769	1852	Ireland	military	politics
12	299	Douglas Macarthur	1880	1964	US	military	military
13	316	Pervez Musharraf	1943		Pakistan	politics	military
14	325	Nawaz Sharif	1949		Pakistan	politics	military
15	334	Ion Antonescu	1882	1946	Romania	military	lit
16	340	Frederick II of Prussia	1712	1786	Germany	nobility	military
17	341	Oliver Cromwell	1599	1658	England	military	politics
18	378	Simón Bolívar	1783	1830	Venezuela	military	politics
19	393	Hosni Mubarak	1928		Egypt	military	politics
20	418	William Henry Harrison	1773	1841	US	politics	military
		Top decile (random sample)					
5737	128776	Franco Lucchini	1917	1943	Italy	military	military
5738	128786	Paul Behncke	1869	1937	Germany	military	military
5739	128795	Thomas Ewing, Jr.	1829	1896	US	law	military
5740	128826	Angelo d 27arrigo	1961	2006	Italy	military	education
5741	128829	Amédée Mouchez	1821	1892	France	military	military
		First quartile (random sample)					
14344	324354	Friedrich Von Hollmann	1842	1913	Germany	military	nobility
14345	324380	Focko Ukena	1360	1435	.	military	politics
14347	324442	Eric Plant	1890	1950	Australia	military	military
14346	324501	Toto Koopman	1908	1991	Indonesia	military	arts
14348	324515	Pierre Segrétain	1909	1950	France	military	military
		Median (random sample)					
28689	631898	Edward Everett Smith	1861	1931	US	law	military
28690	631971	Francois Baby (politician)	1768	1852	Canada	military	politics
28691	632121	Mohammad Hossein Jalali			Iran	military	politics
28692	632126	James Bruce	1732	1791	Russia	Family	military
28693	632306	Joseph Reynolds (congressman)	1785	1864	US	business	military
		Third quartile (random sample)					
43034	902733	Ira Jones	1923	2004	.	lit	military
43035	902735	Paterson Fraser	1907	2001	.	military	military
43037	902766	Kent Foster	1937		Canada	military	military
43038	902802	Cornelius Coffey	1903	1994	US	military	education
43036	902961	Earle D. Chesney	1900	1966	.	religious	military
		Last decile (random sample)					
51644	1071958	Willoughby Williams		1802	US	military	politics
51642	1072099	George M. Cox	1892	1977	.	military	lit
51645	1072143	Richard O 27Farrell		1757	.	lit	military
51643	1072323	William Murray Threipland	1866	1942	England	military	military
51641	1072475	Arthur Alphin			US	military	military

Table A.9: Military

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	52	Steve Jobs	1955	2011	US	business	inventor
2	57	Leonardo da Vinci	1452	1519	Italy	inventor	arts
3	84	Christopher Columbus	1451	1506	Italy	inventor	inventor
4	89	Nikola Tesla	1856	1943	Austria	inventor	business
5	134	Thomas Edison	1847	1931	US	inventor	business
6	160	Alexander Graham Bell	1847	1922	England	studies	inventor
7	255	Hernán Cortés	1485	1547	Spain	inventor	nobility
8	285	James Cook	1728	1779	England	inventor	inventor
9	404	Ibn Battuta	1304	1369	.	inventor	lit
10	410	James Cameron	1954		Canada	arts	inventor
11	449	Edmund Hillary	1919	2008	New Zealand	sports	inventor
12	632	Ferdinand Magellan	1480	1521	Portugal	inventor	inventor
13	633	Fridtjof Nansen	1861	1930	Norway	inventor	studies
14	733	Alfred Russel Wallace	1823	1913	England	studies	inventor
15	754	Roald Amundsen	1872	1928	Norway	inventor	politics
16	760	Guglielmo Marconi	1874	1937	Italy	inventor	business
17	910	James Watt	1736	1819	Scotland	inventor	business
18	1016	Dmitri Mendeleev	1834	1907	Russia	studies	inventor
19	1311	Tim Berners-Lee	1955		England	studies	inventor
20	1438	Giacomo Casanova	1725	1798	Italy	inventor	lit
			Top decile (random sample)				
931	90515	Sámuel Teleki	1845	1916	Hungary	inventor	inventor
932	90576	Jacob Aaron Westervelt	1800	1879	.	inventor	lit
933	90597	Otto Schmitt	1913	1998	US	inventor	business
934	90617	Royal Rife	1888	1971	US	inventor	arts
935	91126	Jozef Murgaš	1864	1929	Slovakia	inventor	arts
			First quartile (random sample)				
2331	255786	Aaron Seigo	1975		Canada	business	inventor
2332	256008	David Wilhelm	1956		.	inventor	politics
2333	256020	Gerard Unger	1942		Netherlands	arts	inventor
2334	256092	Michael J. Freeman	1947		US	inventor	business
			Median (random sample)				
4663	580149	Pierre de Sales Laterrière	1740	1815	France	inventor	politics
4664	580213	Alexis Nihon	1902	1980	Belgium	inventor	business
4665	580260	Richard K. Diran	1949		US	inventor	arts
4666	580468	James H. Kelley	1833	1912	US	inventor	business
			Third quartile (random sample)				
6994	890906	Dudley Bradstreet	1711	1763	Ireland	inventor	politics
6995	891735	Ira P. DeLoache	1879	1965	US	business	inventor
6996	892276	Wilhelm Brenneke	1865	1951	Germany	inventor	inventor
6997	892363	Silas C. Overpack	1842	1918	.	business	inventor
			Last decile (random sample)				
8394	1087816	Eric Drew			.	inventor	lit
8395	1088485	Jerome Wheelock	1834	1902	US	inventor	inventor
8396	1088951	Stephen White (programmer)	1969		Canada	business	inventor
8398	1089168	A. J. R. Russell-Wood	1940	2010	Brazil	studies	inventor
8397	1089289	Nellie Zabel Willhite	1892	1991	US	inventor	Other

Table A.10: Inventors

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	3	Jesus	0	30	.	Family	religious
2	13	Mahomet	570	632	.	religious	Other
3	15	Pope Francis	1936		Argentina	religious	religious
4	17	Pope John Paul Ii	1920	2005	Poland	politics	religious
5	33	Pope Benedict Xvi	1927		Germany	politics	religious
6	46	Martin Luther	1483	1546	Germany	religious	education
7	77	Søren Kierkegaard	1813	1855	Denmark	studies	religious
8	100	Augustine of Hippo	354	430	.	religious	studies
9	112	Thomas Aquinas	1225	1274	Italy	religious	religious
10	115	Mary (mother of Jesus)	-18	41	.	religious	Family
11	169	14th Dalai Lama	1935		.	religious	studies
12	238	Joseph Smith	1805	1844	US	religious	politics
13	272	Mother Teresa	1910	1997	.	religious	religious
14	321	Ruhollah Khomeini	1902	1989	Iran	religious	politics
15	351	Umar Farooq	588	644	.	religious	law
16	376	Joan of Arc	1412	1431	France	religious	nobility
17	380	Abdullah of Saudi Arabia	1924	2015	Saudi Arabia		religious
18	385	Dante Alighieri	1265	1321	Italy	lit	religious
19	422	Pope John Xxiii	1881	1963	Italian	Family	religious
20	440	Anselm of Canterbury	1033	1109	.	religious	religious
			Top decile (random sample)				
4698	118972	José Correia da Serra	1750	1823	Portugal	religious	studies
4699	118996	Tadeusz Isakowicz-Zaleski	1956		Poland	religious	lit
4700	119042	Valerio Valeri	1883	1963	Italy	religious	religious
4702	119133	Mahasena of Anuradhapura		301	Sri Lanka	nobility	religious
4701	119136	Yehuda Alharizi	1165	1225	Spain	religious	studies
			First quartile (random sample)				
11748	285318	Franz Ludwig Von Erthal	1730	1795	.	religious	religious
11749	285324	Carlo Domenico del Carretto	1454	1514	Italy	religious	religious
11752	285366	Sam Pollard	1864	1915	England	religious	lit
11751	285377	Leo Soekoto	1920	1995	.	religious	religious
11750	285382	Antonius Maria Bodewig	1839	1915	Germany	religious	Other
			Median (random sample)				
23499	590055	Brian Hennessy (bishop)	1919	1997	US	religious	religious
23498	590072	Nick Ribush			.	religious	Other
23501	590103	Soulmother of Küssnacht		1577	Switzerland	Other	religious
23500	590229	James Bilsborrow	1862	1931	England	religious	religious
			Third quartile (random sample)				
35249	893066	Harold Hyde-Lees	1890	1963	England	religious	religious
35251	893091	Hyacinth (Jacek) Gulski	1847	1911	Poland	business	religious
35250	893134	Keshavananda Brahmachari		1942	India	religious	religious
35248	893220	David Saperstein (rabbi)	1947		US	religious	law
35247	893224	Henry Marshall (bishop of Salford)	1884	1955	Italy	religious	religious
			Last decile (random sample)				
42300	1084730	Deborah VanAmerongen			US	politics	religious
42297	1084790	George E. Hibbard	1924	1991	US	religious	arts
42301	1084799	Erasmus Stourton	1603	1658	.	religious	Other
42298	1084862	Philip Remler			US	politics	religious
42299	1084911	Thomas Bache (judge)		1410	Italy	religious	law

Table A.11: Religion

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	121	Confucius	-550	-470	.	education	lit
2	247	Isaac Asimov	1920	1992	US	lit	education
3	496	Martin Heidegger	1889	1976	Germany	studies	education
4	592	J. Robert Oppenheimer	1904	1967	US	studies	education
5	599	Karl Popper	1902	1994	England	studies	education
6	686	Pierre-Simon Laplace	1749	1827	France	education	politics
7	793	Talcott Parsons	1902	1979	US	studies	education
8	795	Emily Dickinson	1830	1886	US	lit	education
9	894	Paul Krugman	1953		US	studies	education
10	1024	Joseph Stiglitz	1943		US	studies	education
11	1065	Ralph Waldo Emerson	1803	1882	US	lit	education
12	1101	Euripides	-480	-400	Greece	lit	education
13	1138	Heraclitus	-530	-470	Greece	studies	education
14	1207	Golda Meir	1898	1978	Israel	education	politics
15	1216	Ibn Taymiyyah	1263	1328	.	education	religious
16	1264	Robert H. Goddard	1882	1945	US	business	education
17	1493	Robert Baden-Powell, 1st Baron Baden-Powell	1857	1941	.	education	politics
18	1556	Origen	185	254	.	education	religious
19	1558	Aaron Copland	1900	1990	US	arts	education
20	1724	Petrarch	1304	1374	Italy	education	lit
		Top decile (random sample)					
6776	223037	Michael Jensen	1939		US	education	education
6777	223149	Robert B. Wilson	1937		US	studies	education
6778	223153	Antoine-Jean Saint-Martin	1791	1832	France	education	studies
6779	223202	Anne Morelli	1948		Belgium	studies	education
6780	223239	Sa'id Akhtar Rizvi	1927	2002	Tanzania	education	lit
		First quartile (random sample)					
16944	444491	Ida Bieler			US	arts	education
16942	444520	Jacob Chandy	1910	2007	India	studies	education
16943	444530	Arthur W. Barton	1899	1976	.	business	education
16945	444552	Konstantin Posse	1847	1928	Russia	studies	education
		Median (random sample)					
33886	741355	Michael Stumpf	1970		.	education	studies
33888	741395	George Kennion	1845	1922	England	education	education
33885	741481	Floyd Graham	1902	1974	US	arts	education
33887	741538	Gong Xiantian	1944		.	education	education
		Third quartile (random sample)					
50831	971470	Thomas H. Makiyama	1928	2005	US	sports	education
50828	971554	Kerreen Reiger			Australia	education	lit
50829	971607	Keith Vivian Alexander			New Zealand	education	business
50830	971716	Thomas Chase (educator)	1827	1892	US	education	education
		Last decile (random sample)					
60997	1116884	Víctor Manuel Gutiérrez	1922	1966	Guatemala	education	politics
60993	1116955	El-Gorashy	1942	1964	.	education	education
60994	1117025	Sig Andrusking	1913	1994	US	sports	education
60996	1117251	Achsah Guibbory			US	education	education
60995	1117294	Barry Evans (rugby union)	1962		England	sports	education

Table A.12: Education

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	18	Alexander The Great	-350	-320	Greece	nobility	nobility
2	60	Charlemagne	740	814	.	nobility	nobility
3	69	Elizabeth Ii	1926		England	nobility	
4	88	Louis XIV of France	1638	1715	France	nobility	nobility
5	119	Elizabeth I of England	1533	1603	England	nobility	Family
6	122	Diana, Princess of Wales	1961	1997	England	Family	nobility
7	125	Constantine The Great	272	337	Italy	nobility	Family
8	128	Genghis Khan	1160	1227	Mongolia	Other	nobility
9	143	Akbar	1542	1605	.	nobility	nobility
10	161	Queen Victoria	1819	1901	England	nobility	nobility
11	182	Nicholas II of Russia	1868	1918	Russia	nobility	nobility
12	194	Charles, Prince of Wales	1948		England	Family	nobility
13	196	Henry VIII of England	1491	1547	England		nobility
14	200	Muawiyah I	602	680	Syria	nobility	nobility
15	209	Trajan	53	117	Italy	nobility	politics
16	215	Haile Selassie	1892	1975	Ethiopia	nobility	nobility
17	255	Hernán Cortés	1485	1547	Spain	inventor	nobility
18	258	Nero	37	68	Italy	nobility	nobility
19	278	Hadrian	76	138	Italy	nobility	inventor
20	292	Marcus Aurelius	121	180	Italy	nobility	studies
		Top decile (random sample)					
3564	36235	Euthydemus I		-200	.	nobility	nobility
3565	36268	Ivan Stephen of Bulgaria		1373	Serbia	Family	nobility
3566	36271	Kakinomoto No Hitomaro	662	710	Japan	lit	nobility
3567	36283	Richard Woodville, 1st Earl Rivers	1405	1469	England	nobility	nobility
3568	36300	George, Duke of Bavaria	1455	1503	Germany	nobility	Family
		First quartile (random sample)					
8913	128857	Louis de Silvestre	1675	1760	France	arts	nobility
8914	128859	Turan-Shah		1180	Turkey	nobility	nobility
8915	128869	Fergus of Galloway		1161	Scotland	nobility	nobility
8917	128914	Li Lianying	1848	1911	China	nobility	nobility
8916	128887	Alexander Mourousis		1816	Moldova	nobility	nobility
		Median (random sample)					
17830	356840	Isabelle Romée	1377	1458	.	Family	nobility
17829	356780	Tsuneharu Takeda	1944		Japan	politics	nobility
17828	356866	Werner Ewald	1914	1993	Germany	Other	nobility
17830	356932	Kim Myeong-Won	1534	1602	.	politics	nobility
17831	356909	Antoni Jan Ostrowski	1782	1845	Poland	nobility	business
		Third quartile (random sample)					
26743	716750	William St Clair of Roslin		1778	.	nobility	nobility
26744	716907	Giovanni dalle Carceri		1358	.	nobility	nobility
26745	716936	Robert Logan of Restalrig	1550	1606	Scotland	nobility	nobility
26746	716956	Ludwig Spindler	1910	1944	Germany	military	nobility
26747	716971	Thihathura II of Ava	1474	1501	.	nobility	nobility
		Last decile (random sample)					
32092	1026091	James Neuberger	1949		.	politics	nobility
32094	1026099	Abdulaziz Mohammed Majid Al-Farsi	1976		.	lit	nobility
32095	1026134	Rick Nevin			US	politics	nobility
32093	1026249	Robert III de Stuteville		1186	England	nobility	nobility
32096	1026616	Thomas Farrer, 2nd Baron Farrer	1859	1940	.	nobility	Family

Table A.13: Nobility

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	65	Thomas Jefferson	1743	1826	US	law	politics
2	167	John Adams	1735	1826	US	law	lit
3	280	William Howard Taft	1857	1930	US	law	politics
4	351	Umar Farooq	588	644	.	religious	law
5	369	Chester A. Arthur	1829	1886	US	law	politics
6	487	Thomas More	1478	1535	England	law	studies
7	629	Muhammad Ali Jinnah	1876	1948	Pakistan	law	politics
8	634	Al-Ghazali	1058	1111	.	religious	law
9	641	Michelle Obama	1964		US	law	lit
10	670	Maximilien de Robespierre	1758	1794	France	law	politics
11	767	Hassan Rouhani	1948		Iran	politics	law
12	908	Muhammad Zia-Ul-Haq	1924	1988	Pakistan	politics	law
13	925	Ehud Olmert	1945		Israel	politics	law
14	1043	Jeremy Bentham	1748	1832	England	studies	law
15	1239	Rudy Giuliani	1944		US	law	business
16	1257	Rick Santorum	1958		US	law	politics
17	1540	Mohamed Elbaradei	1942		Egypt	law	education
18	1573	Montesquieu	1689	1755	France	law	lit
19	1636	Alben W. Barkley	1877	1956	US	law	politics
20	1703	Sonia Sotomayor	1954		US	law	law
			Top decile (random sample)				
4330	258340	Ramón J. Cárcano	1860	1946	Argentina	law	studies
4331	258402	András Tasnádi Nagy	1882	1956	Hungary	politics	law
4332	258418	Malcolm McCusker	1938		Australia	law	Other
4333	258440	Lloyd L. Gaines	1911		Canada	law	politics
4334	258475	John Trevor (speaker)	1630	1717	Wales	law	politics
			First quartile (random sample)				
10829	522365	Patrick Redmond	1966		England	lit	law
10830	522398	Jerome J. Shestack	1923	2011	US	law	law
10828	522442	Hippolyte Rolin	1804	1883	Belgium	law	politics
10831	522539	Filippo Mancuso	1922	2011	Italy	law	politics
10832	522549	Amédée Dunois	1878	1945	France	law	lit
			Median (random sample)				
21657	814976	Martin S. Ackerman	1932	1993	.	law	business
21661	815002	James M. Burns (judge)	1924	2001	US	law	law
21659	815010	William Travers Jerome	1859	1934	US	law	politics
21658	815099	Robert H. Johnson	1916	2011	US	lit	law
21660	815158	A. David Mazzone	1928	2004	US	law	law
			Third quartile (random sample)				
32489	1040410	Ian West (Australian politician)	1951		Australia	politics	law
32486	1040427	Andrew Bridge (lawyer)			US	law	politics
32490	1040629	Tony Randerson			New Zealand	military	law
32487	1040718	John Batiuk	1923	2005	Canada	politics	law
32488	1040933	William Cliffe		1558	England	religious	law
			Last decile (random sample)				
38984	1160517	Des Adam	1945		.	business	law
38985	1160793	Charles A. Cooke	1848	1917	US	politics	law
38987	1160914	William A. Denning	1817	1856	US	law	politics
38986	1160918	Lefteris Zagoritis	1956		Greece	law	politics
38988	1161388	T. Veeraswamy			India	politics	law

Table A.14: Law

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	29	Hillary Rodham Clinton	1947		US	politics	politics
2	55	Margaret Thatcher	1925	2013	England	politics	politics
3	59	Lady Gaga	1986		US	arts	arts
4	69	Elizabeth Ii	1926		England	nobility	
5	72	Serena Williams	1981		US	sports	sports
6	78	Celine Dion	1968		Canada	arts	arts
7	80	Madonna (entertainer)	1958		US	arts	arts
8	87	Marilyn Monroe	1926	1962	US	arts	arts
9	94	Whitney Houston	1963	2012	US	arts	arts
10	96	Beyoncé Knowles	1981		US	arts	arts
11	98	Taylor Swift	1989		US	arts	arts
12	103	Angela Merkel	1954		Germany	politics	studies
13	113	Meryl Streep	1949		US	arts	arts
14	115	Mary (mother of Jesus)	-18	41	.	religious	Family
15	116	Yulia Tymoshenko	1960		Ukraine	politics	politics
16	119	Elizabeth I of England	1533	1603	England	nobility	Family
17	122	Diana, Princess of Wales	1961	1997	England	Family	nobility
18	123	Maria Sharapova	1987		.	sports	sports
19	129	Benazir Bhutto	1953	2007	Pakistan	politics	politics
20	132	Angelina Jolie	1975		US	arts	arts
		Top decile (random sample)					
19519	115708	Laura Riding	1901	1991	US	lit	lit
19522	115715	Alice Rohrwacher	1982		Italy	arts	lit
19521	115719	Jan Lehane	1941		Australia	sports	sports
19520	115722	Fleur East	1987		England	arts	arts
19523	115728	Tatjana Patitz	1966		Germany	arts	arts
		First quartile (random sample)					
48804	301638	Josephine Meckseper	1964		Germany	arts	arts
48805	301644	Livia Zita	1984		Hungary	arts	arts
48802	301645	Anne Rosellini			US	arts	arts
48801	301668	Brunilde Sismondo Ridgway	1929		Italy	studies	arts
48803	301716	Griedge Mbock Bathy	1995		France	sports	sports
		Median (random sample)					
97603	611748	Nancy Buchanan	1946		US	arts	arts
97607	611753	Kelly Cassidy			US	politics	politics
97604	611756	Suzie Faulkner	1979		Australia	sports	sports
97606	611763	Kaitlin Cochran	1987		US	sports	sports
97605	611778	Merle Goldman	1931		US	studies	education
		Third quartile (random sample)					
146406	913747	Marion Goldman			US	education	religious
146409	913916	Lila Rose Kaplan	1980		US	lit	arts
146408	913923	Suzanne Goldenberg	1962		Canada	lit	lit
146405	913949	Claire Keelan			England	arts	arts
146407	914089	María Dolores 22Mary 22 Tarrero-Serrano	1924	2010	Cuba	politics	Family
		Last decile (random sample)					
175688	1109637	Claudia Dain			US	lit	lit
175689	1110020	Oksana Ryabinicheva	1990		Russia	sports	sports
175690	1109848	Mary Kay (landscape photographer)			Greece	lit	arts
175691	1110065	Monica Buck			US	Other	Other
175692	1110086	Valentyna Brik	1985		Ukraine	sports	sports

Table A.15: Women, all categories

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
			Top 20 individuals				
1	29	Hillary Rodham Clinton	1947		US	politics	politics
2	55	Margaret Thatcher	1925	2013	England	politics	politics
3	69	Elizabeth Ii	1926		England	nobility	
4	103	Angela Merkel	1954		Germany	politics	studies
5	115	Mary (mother of Jesus)	-18	41	.	religious	Family
6	116	Yulia Tymoshenko	1960		Ukraine	politics	politics
7	119	Elizabeth I of England	1533	1603	England	nobility	Family
8	122	Diana, Princess of Wales	1961	1997	England	Family	nobility
9	129	Benazir Bhutto	1953	2007	Pakistan	politics	politics
10	161	Queen Victoria	1819	1901	England	nobility	nobility
11	212	Sarah Palin	1964		US	politics	lit
12	231	Catherine The Great	1729	1796	Russia	politics	Other
13	242	Indira Gandhi	1917	1984	India	politics	politics
14	259	Aung San Suu Kyi	1945		Myanmar	politics	politics
15	272	Mother Teresa	1910	1997	.	religious	religious
16	276	Corazon Aquino	1933	2009	.	politics	lit
17	298	Marine Le Pen	1968		France	politics	politics
18	306	Katharine Hepburn	1907	2003	US	arts	politics
19	322	Dilma Rousseff	1947		Brazil	studies	politics
20	356	Audrey Hepburn	1929	1993	England	arts	politics
			Top decile (random sample)				
4048	119361	Katrín Jakobsdóttir	1976		Iceland	politics	politics
4049	119535	Yordanka Fandakova	1962		Bulgaria	politics	politics
4050	119610	Diana E. H. Russell	1938		South Africa	politics	lit
4051	119618	Jacqueline Auriol	1917	2000	France	military	military
4052	119619	Vera Kobalia	1981		Georgia	politics	politics
			First quartile (random sample)				
10124	338478	Brigitte Douay	1947		France	politics	politics
10123	338502	Christine Boyer	1771	1800	.	Family	religious
10126	338556	Linda Jeffrey	1958		Canada	politics	politics
10125	338624	Mary Kim Titla	1960		US	lit	law
10127	338672	Veronica Palm	1973		Sweden	politics	politics
			Median (random sample)				
20247	661532	Karen Tandy			US	Other	law
20251	661538	Sally Huffer	1965		US	politics	politics
20248	661549	Helene Moszkiewiez	1920	1998	.	military	politics
20250	661585	Maria Pilar Riba Font	1944		Spain	politics	politics
20249	661613	Tané Matsukata	1918	1989	Japan	Other	politics
			Third quartile (random sample)				
30372	931559	Caroline Keer	1857	1928	England	military	studies
30371	931577	Anna Walker (civil servant)	1951		England	politics	politics
30373	931612	Mina Adampour	1987		Norway	lit	politics
30374	931680	Rosamond Carr	1912	2006	US	politics	lit
30375	932156	Romola Sinha	1913	2010	India	politics	politics
			Last decile (random sample)				
36450	1133938	Angie Mentink	1973		.	sports	military
36446	1134072	Kim White			US	politics	business
36448	1134362	Majaji		350	South Africa	politics	nobility
36449	1134369	Danielle Moore	1946		US	education	politics
36447	1134454	Mariam Mfaki	1946	2015	Tanzania	politics	politics

Table A.16: Women in governance

Cat. rank	Ov. rank	Name	Birth	Death	Citizenship	Occupation B1	Occupation B2
		Top 20 individuals					
1	5	Napoleon Bonaparte	1769	1821	France	military	politics
2	28	Charles de Gaulle	1890	1970	France	lit	politics
3	88	Louis XIV of France	1638	1715	France	nobility	nobility
4	90	Voltaire	1694	1778	France	lit	studies
5	111	Nicolas Sarkozy	1955		France	politics	politics
6	138	Napoleon Iii	1808	1873	France	politics	politics
7	239	Victor Hugo	1802	1885	France	lit	lit
8	244	René Descartes	1596	1650	France	studies	studies
9	260	Paul Gauguin	1848	1903	France	arts	arts
10	298	Marine Le Pen	1968		France	politics	politics
11	301	Zinedine Zidane	1972		France	sports	sports
12	308	Jules Verne	1828	1905	France	lit	lit
13	310	Louis XVI of France	1754	1793	France	nobility	Family
14	370	François Mitterrand	1916	1996	France	politics	politics
15	376	Joan of Arc	1412	1431	France	religious	nobility
16	402	François Hollande	1954		France	politics	politics
17	413	Louis Pasteur	1822	1895	France	studies	studies
18	433	Jacques Chirac	1932		France	politics	politics
19	434	Franck Ribéry	1983		France	sports	sports
20	448	Jean-Paul Sartre	1905	1980	France	studies	lit
		Top decile (random sample)					
3918	77836	Jean Chouan	1757	1794	France	religious	politics
3919	77842	Vladimir Solomonovich Pozner	1905	1992	France	lit	studies
3920	77865	Henri Joseph Anastase Perrotin	1845	1904	France		
3921	77881	Raymond Delisle	1943	2013	France	sports	sports
3922	77911	Marcel Griaule	1898	1956	France	studies	studies
		First quartile (random sample)					
9798	188494	Bernard Borderie	1924	1978	France	arts	lit
9800	188593	Teddy da Costa	1986		France	sports	sports
9799	188594	Gilbert Sinoué	1947		France	arts	lit
9801	188603	Alexandre Guiraud	1788	1847	France	lit	arts
9802	188641	Félicien Menu de Ménil	1860	1930	France	arts	arts
		Median (random sample)					
19601	378843	Raymond Abad	1930		France	sports	sports
19600	378857	Auguste Mermet	1810	1889	France	arts	arts
19598	378867	Bruno Étienne	1937	2009	France	studies	politics
19599	378939	Amédée Pichot	1795	1877	France	studies	studies
		Third quartile (random sample)					
29397	622687	Jean-Baptiste Joseph Émile Montégut	1825	1895	France	lit	lit
29398	622819	Georges William Thornley	1857	1935	France	arts	arts
29399	622876	Philippe Berre	1954		France	business	arts
29400	622932	Hippolyte Laroche	1848	1914	France	military	inventor
29401	622990	Joseph Caillot	1733	1816	France	arts	arts
		Last decile (random sample)					
35277	888700	Ivan Grésèque			France	sports	sports
35279	888966	Julien Bègue	1993		France	sports	sports
35280	889175	Michael Clarke (priest)	1935	1978	France	religious	education
35281	889233	Henri Monteux	1874	1943	France	arts	arts
35278	889262	Nicolas Ladvocat-Billiard		1681	France	religious	studies

Table A.17: French individuals

A.3 `Wikipedia` Page - Ray Charles

Ray Charles

From Wikipedia, the free encyclopedia

Life and career [edit]

1930–45: Early years [edit]

Ray Charles Robinson was the son of Aretha (née William) Robinson,[15] a sharecropper, and Bailey Robinson, a railroad repair man, mechanic, and handyman.[16] When Charles was an infant, his family moved from his birthplace in Albany, Georgia back to his mother's hometown of Greenville, Florida.

Charles had little contact with his father growing up, and it is unclear whether his mother and father were ever married. Charles was raised by his biological mother Aretha, as well as his father's first wife, a woman named Mary Jane. Growing up, he referred to Aretha as "Mama", and Mary Jane as "mother".[11] Aretha was a devout Christian, and the family attended the New Shiloh Baptist Church.[15]

In his early years, Charles showed a curiosity for mechanical objects, and would often watch his neighbors working on their cars and farm machinery. His musical curiosity was sparked at Mr. Wylie Pitman's Red Wing Cafe, when Pitman played boogie woogie on an old upright piano; Pitman subsequently taught Charles how to play piano himself. Charles and his mother were always welcome at the Red Wing Cafe, and even lived there when they were experiencing financial difficulties.[11] Pitman would also care for Ray's brother George, to take the burden off Aretha. George drowned in Aretha's laundry tub when he was four years old, and Ray was five.[11][16] Charles started to lose his sight at the age of four[3] or five,[17] and was completely blind by the age of seven, apparently as a result of glaucoma.[18] Broke, uneducated and still mourning the loss of Charles' brother George, Aretha used her connections in the local community to find a school that would accept blind African American students. Despite his initial protest, Charles would attend school at the Florida School for the Deaf and the Blind in St. Augustine from 1937 to 1945.[19]

Charles began to develop his musical talent at school,[18] and was taught to play the classical piano music of Bach, Mozart and Beethoven. His teacher Mrs. Lawrence taught him how to read music using braille, a difficult process that requires learning the left hand movements by reading braille with the right hand and learning the right hand movements by reading braille with the left hand, and then synthesizing the two parts. While Charles was happy to play the piano, he was more interested in the jazz and blues music he heard on the family radio than classical music.[19] On Fridays, the South Campus Literary Society held assemblies where Charles would play piano and sing popular songs. On Halloween and Washington's birthday, the black Department of the school had socials where Charles would play. It was here he established "RC Robinson and the Shop Boys" and sang his own arrangement of "Jingle Bell Boogie". During this time, he performed on WFOY radio in St. Augustine.[19]

Ray Charles statue in Greenville, Florida

Aretha died in the spring of 1945, when Charles was 14 years old. Her death came as a shock to Ray, who would later consider the deaths of his brother and mother to be "the two great tragedies" of his life. Charles returned to school after the funeral, but was then expelled in October for playing a prank on his teacher.[19]

1945–52: Life in Florida, Los Angeles, Seattle and first hits [edit]

After leaving school, Charles moved to Jacksonville with a couple who were friends of his mother. He played the piano for bands at the Ritz Theatre in LaVilla for over a year, earning $4 a night. He also joined the musicians' union in the hope that it would help him get work. He befriended many union members, but others were less kind to him because he would monopolize the union hall's piano, since he did not have one at home. He started to build a reputation as a talented musician in Jacksonville, but the jobs did not come fast enough for him to construct a strong identity. He decided to leave Jacksonville and move to a bigger city with more opportunities.[20]

At age 16, Charles moved to Orlando, where he lived in borderline poverty and went without food for days. It was an extremely difficult time for musicians to find work, as since World War II had ended there were no "G.I. Joes" left to entertain. Charles eventually started to write arrangements for a pop music band, and in the summer of 1947 he unsuccessfully auditioned to play piano for Lucky Millinder and his sixteen-piece band.[19]

In 1947, Charles moved to Tampa, where he had two jobs: one as a pianist for Charlie Brantley's Honeydippers,[21] a seven-piece band; and another as a member of a white country band called The Florida Playboys (though there is no historical trace of Charles' involvement in The Florida Playboys besides Charles' own testimony). This is where he began his habit of always wearing sunglasses, made by designer Billy Stickles. In his early career, he modeled himself on Nat "King" Cole. His first four recordings—"Wondering and Wondering", "Walking and Talking", "Why Did You Go?" and "I Found My Baby There"—were supposedly made in Tampa, although some discographies also claim he recorded them in Miami in 1951, or Los Angeles in 1952.[19]

Charles had always played piano for other people, but he was keen to have his own band. He decided to leave Florida for a large city, and, considering Chicago and New York City too big, followed his friend Gossie McKee to Seattle, Washington in March 1948, knowing that the biggest radio hits came from northern cities.[18][22] Here he met and befriended, under the tutelage of Robert Blackwell, a 15-year-old Quincy Jones.[23]

He started playing the one-to-five A.M. shift at the Rocking Chair with his band McSon Trio, which featured McKee on guitar and Milton Garrett on bass. Publicity photos of the trio are some of the earliest recorded photographs of Ray Charles. In April 1949, Charles and his band recorded "Confession Blues", which became his first national hit, soaring to the second spot on the Billboard R&B chart.[22] While still working at the Rocking Chair, he also arranged songs for other artists, including Cole Porter's "Ghost of a Chance" and Dizzy Gillespie's "Emanon".[20] After the success of his first two singles, Charles moved to Los Angeles in 1950, and spent the next few years touring with blues artist Lowell Fulson as his musical director.[3]

In 1950, his performance in a Miami hotel would impress Henry Stone, who went on to record a Ray Charles Rockin' record (which never became particularly popular). During his stay in Miami, Charles was required to stay in the segregated but thriving black community of Overtown. Stone later helped Jerry Wexler find Charles in St. Petersburg.[24]

After joining Swing Time Records, he recorded two more R&B hits under the name "Ray Charles": "Baby, Let Me Hold Your Hand" (1951), which reached number five; and "Kissa Me Baby"(1952), which reached number eight. Swing Time folded the following year, and Ahmet Ertegün signed him to Atlantic Records.[18]

1952–59: Signing with Atlantic Records [edit]

This section **relies largely or entirely upon a single source**. Relevant discussion may be found on the talk page. Please help improve this article by introducing citations to additional sources. *(April 2015)*

In June 1952, Atlantic Records bought Ray's contract for $2,500.[25] Charles' first recording session with Atlantic ("The Midnight Hour"/"Roll With my Baby") took place in September 1952, although his last Swingtime release ("Misery in my Heart"/"The Snow is Falling") would not appear until February 1953. He began recording jump blues and boogie-woogie style recordings as well as slower blues ballads, in which he continued to show the vocal influences of Nat "King" Cole and Charles Brown. "Mess Around" became Charles' first Atlantic hit in 1953; the following year he had hits with "It Should Have Been Me" and "Don't You Know," which became his first chart success for Atlantic.[25] He also recorded the songs "Midnight Hour" and "Sinner's Prayer." Some elements of his own vocal style were evident in "Sinner's Prayer," "Mess Around," and "Don't You Know".[citation needed]

Late in 1954, Charles recorded his own composition "I Got a Woman"; the song became one of his most notable hits, reached number two on the R&B chart.[25] "I Got a Woman" included a mixture of gospel, jazz and blues elements that would later prove to be seminal in the development of rock 'n' roll and soul music. In 1955, he had hits "This Little Girl of Mine" and "A Fool for You". In upcoming years, he scored with "I'll Drown in My Own Tears" and "Hallelujah, I Love Her So." By 1959, Ray Charles reached the *Billboard Top Ten* with "What'd I Say" which made him a major figure in R&B.[25]

Parallel to his R&B career, Charles also recorded instrumental jazz albums such as 1957's The Great Ray Charles. During this time, Charles also worked with jazz vibraphonist Milt Jackson, releasing *Soul Brothers* in 1958 and *Soul Meeting* in 1961. By 1958, Charles was not only headlining black venues such as The Apollo Theater and The Uptown Theater, but also bigger venues such as The Newport Jazz Festival (where he would cut his first live album). In 1956, Charles recruited a young all-female singing group named the Cookies, and reshaped them as The Raeletes. Up to this point, Charles had used his wife and other musicians to back him on recordings such as "This Little Girl of Mine" and "Drown In My Own Tears". The Raelettes' first recording session with Charles was on the bluesy-gospel inflected "Leave My Woman Alone".[citation needed]

1959–67: Crossover success [edit]

See also: What'd I Say and Modern Sounds in Country and Western Music

Charles reached the pinnacle of his success at Atlantic with the release of "What'd I Say", a complex song that combined gospel, jazz, blues and Latin music, which Charles would later claim he had composed spontaneously as he was performing in clubs and dances with his small band. Despite some radio stations banning the song because of its sexually suggestive lyrics, the song became Charles' first ever crossover top ten pop record.[26] Later in 1959, he released his first country song (a cover of Hank Snow's "I'm Movin' On"), as well as recording three more albums for the label: a jazz record (later released in 1961 as *The Genius After Hours*); a blues record (released in 1961 as *The Genius Sings the Blues*); and a traditional pop/big band record (*The Genius of Ray Charles*). *The Genius of Ray Charles* provided his first top 40 album entry, where it peaked at No. 17, and was later held as a landmark record in Charles' career.[citation needed]

Charles in 1971. Photo: Heinrich Klaffs.

Charles' Atlantic contract expired in the fall of 1959, with several big labels offered him record deals; choosing not to renegotiate his contract with Atlantic, Ray Charles signed with ABC-Paramount Records in November 1959.[27] He obtained a much more liberal contract than other artists had at the time, with ABC offering him a $50,000 annual advance, higher royalties than before and eventual ownership of his masters—a very valuable and lucrative deal at the time.[28] During his Atlantic years, Charles had been heralded for his own inventive compositions, but by the time of the release of the instrumental jazz LP *Genius + Soul = Jazz* (1960) for ABC's subsidiary label Impulse!, he had virtually given up on writing original material, instead following his eclectic impulses as an interpreter.[26]

With "Georgia on My Mind", his first hit single for ABC-Paramount in 1960, Charles received national acclaim and four Grammy Awards, including two for "Georgia on My Mind": Best Vocal Performance Single Record or Track, Male and Best Performance by a Pop Single Artist. Originally written by composers Stuart Gorrell and Hoagy Carmichael, the song was Charles' first work with Sid Feller, who produced, arranged and conducted the recording.[26][29] Charles earned another Grammy for the follow-up "Hit the Road Jack", written by R&B singer Percy Mayfield.[30]

By late 1961, Charles had expanded his small road ensemble to a full-scale big band, partly as a response to increasing royalties and touring fees, becoming one of the few black artists to crossover into mainstream pop with such a level of creative control.[26][31] This success, however, came to a momentary halt during a concert tour in November 1961, when a police search of Charles' hotel room in Indianapolis, Indiana, led to the discovery of heroin in his medicine cabinet. The case was eventually dropped, as the search lacked a proper warrant by the police, and Charles soon returned to music.[31]

In the early 1960s, whilst on the way from Louisiana to Oklahoma City, Charles faced a near-death experience when the pilot of his plane lost visibility, as snow and his failure to use defroster caused the windshield of the plane to become completely covered in ice. The pilot made a few circles in the air before he was finally able to see through a small part of the windshield and land the plane. Charles placed a spiritual interpretation on the event, claiming that "something or someone which instruments cannot detect" was responsible for creating the small opening in the ice on the windshield which enabled the pilot to land the plane safely.[11]

The 1962 album *Modern Sounds in Country and Western Music*, and its sequel *Modern Sounds in Country and Western Music, Vol. 2*, helped to bring country into the musical mainstream. Charles' version of the Don Gibson song *I Can't Stop Loving You* topped the Pop chart for five weeks, stayed at No. 1 in the R&B chart for ten weeks, and also gave him his only number one record in the UK. In 1962, he founded his own record label, Tangerine Records, which ABC-Paramount promoted and distributed.[32][33] He had major pop hits in 1963 with "Busted" (US No. 4) and *Take These Chains From My Heart* (US No. 8).[citation needed]

In 1965, Charles' career was halted once more after being arrested for a third time for heroin use. He agreed to go to rehab to avoid jail time, and eventually kicked his habit at a clinic in Los Angeles. After spending a year on parole, Charles reappeared in the charts in 1966 with a series of hits composed with the fledgling team of Ashford & Simpson, including the dance number "I Don't Need No Doctor", and "Let's Go Get Stoned", which became his first No. 1 R&B hit in several years. His cover of country artist Buck Owens' "Crying Time" reached No. 6 on the pop chart and helped Charles win a Grammy Award the following March. In 1967, he had a top twenty hit with another ballad, "Here We Go Again".[34]

Death [edit]

In 2003, Charles had successful hip replacement surgery and was originally planning to go back on tour, until he began suffering from other ailments. He died at his home in Beverly Hills, California, on June 10, 2004, surrounded by family and friends,[39][40] as a result of acute liver disease.[3] He was 73 years old. His funeral took place on June 18, 2004, at the First AME Church in Los Angeles, with musical peers such as Little Richard in attendance.[41] B.B. King, Glen Campbell, Stevie Wonder and Wynton Marsalis each played a tribute at Charles' funeral.[42] Charles was interred in the Inglewood Park Cemetery.

One of his last public performances, at the 2003 Montreal International Jazz Festival

His final album, *Genius Loves Company*, was released two months after his death, and consists of duets with various admirers and contemporaries: B.B. King, Van Morrison, Willie Nelson, James Taylor, Gladys Knight, Michael McDonald, Natalie Cole, Elton John, Bonnie Raitt, Diana Krall, Norah Jones and Johnny Mathis. The album won eight Grammy Awards, including Best Pop Vocal Album, Album of the Year, Record of the Year and Best Pop Collaboration with Vocals for "Here We Go Again" with Norah Jones, and Best Gospel Performance for "Heaven Help Us All" with Gladys Knight; he also received nods for his duets with Elton John and B.B. King. The album included a version of Harold Arlen's "Over the Rainbow" sung as a duet with Johnny Mathis, which was played at Charles' memorial service.[42]

Two more posthumous albums were released: *Genius & Friends* (2005), a selection of duets recorded from 1997 to 2004 with artists of Charles' choice, including "Big Bad Love" with Diana Ross; and *Ray Sings, Basie Swings* (2006), which combined archive Ray Charles live vocal performances from the mid-1970s recorded from the concert mixing board with new instrumental tracks specially recorded by the contemporary Count Basie Orchestra and other musicians, to create a "fantasy concert" recording.[citation needed]

Figure A.8: Full Page - Ray Charles

Occupation B1	Frequency	Share
Arts	241,042	32.7
Literature/media	103,421	14.0
Sports	392,163	53.2
Total	736,626	100.0

Table A.18: Share of the different occupations (B1) for occupation Entertainment

Occupation B1	Frequency	Share
Education	36,542	33.0
Studies	74,068	67.0
Total	110,610	100.0

Table A.19: Share of the different occupations (B1) for occupation Academics

A.4 Share of occupations, details

Occupation B1	Frequency	Share
Business	48,974	83.0
Inventor	4,783	8.1
Worker	5,271	8.9
Total	59,028	100.0

Table A.20: Share of the different occupations (B1) for occupation Entrepreneur

Occupation B1	Frequency	Share
Family	11,813	100.0
Total	11,813	100.0

Table A.21: Share of the different occupations (B1) for occupation Family

Occupation B1	Frequency	Share
Law	28,585	10.0
Military	40,516	14.2
Nobility	20,013	7.0
Politics	160,696	56.4
Religious	35,104	12.3
Total	284,914	100.0

Table A.22: Share of the different occupations (B1) for occupation Governance

Occupation B1	Frequency	Share
Other	23,712	100.0
Total	23,712	100.0

Table A.23: Share of the different occupations (B1) for occupation Other

Occupation C1	Frequency	Share
Actor	27,523	11.4
Art	22,858	9.5
Actress	22,414	9.3
Singer	20,697	8.6
Film	20,439	8.5
Painter	17,484	7.3
Music	15,200	6.3
Composer	11,246	4.7
Architect	7,705	3.2
Jazz	5,251	2.2

Table A.24: Share of the different occupations (C1) for occupation Arts

Occupation C1	Frequency	Share
Business	13,632	27.8
Engineer	5,890	12.0
Entrepreneur	3,707	7.6
Pioneer	3,108	6.3
Merchant	2,721	5.6
Farmer	2,431	5.0
Executive	2,147	4.4
Owner	2,068	4.2
Chairman	2,038	4.2
Bank	1,589	3.2

Table A.25: Share of the different occupations (C1) for occupation Business

Occupation C1	Frequency	Share
Professor	12,806	35.0
Scholar	4,776	13.1
College	3,972	10.9
Academic	3,517	9.6
Educator	2,708	7.4
University	2,400	6.6
Teacher	2,346	6.4
Education	1,125	3.1
Dean	956	2.6
Principal	761	2.1

Table A.26: Share of the different occupations (C1) for occupation Education

Occupation C1	Frequency	Share
Son	4,879	41.3
Child	2,920	24.7
Daughter of	1,713	14.5
Wife of	1,203	10.2
Grandson	225	1.9
Mother of	203	1.7
Married to	167	1.4
Marriage	156	1.3
Widow	139	1.2
Only child	86	0.7

Table A.27: Share of the different occupations (C1) for occupation Family

Occupation C1	Frequency	Share
Inventor	1,399	29.2
Explorer	845	17.7
Colonial	794	16.6
Settler	279	5.8
Adventurer	271	5.7
Developer	252	5.3
Discoverer	186	3.9
License	117	2.4
Navigator	104	2.2
Conquistador	91	1.9

Table A.28: Share of the different occupations (C1) for occupation Inventor

Occupation C1	Frequency	Share
Law	12,993	45.5
Judge	5,736	20.1
Attorney	2,977	10.4
Jurist	2,182	7.6
Barrister	1,199	4.2
Legislative	1,017	3.6
Justice	955	3.3
Advocate	582	2.0
Legislator	354	1.2
Magistrate	186	0.7

Table A.29: Share of the different occupations (C1) for occupation Law

Occupation C1	Frequency	Share
Writer	20,004	19.3
Author	15,440	14.9
Poet	12,470	12.1
Journalist	11,277	10.9
Television	10,321	10.0
Novelist	5,300	5.1
Photographer	3,844	3.7
Radio	2,966	2.9
Screenwriter	2,508	2.4
News	2,295	2.2

Table A.30: Share of the different occupations (C1) for occupation Literature/media

Occupation C1	Frequency	Share
Army	6,111	15.1
Officer	5,615	13.9
Chief	4,622	11.4
Soldier	3,990	9.8
Military	3,923	9.7
Navy	2,385	5.9
Commander	1,482	3.7
Marine	1,417	3.5
Air force	1,372	3.4
Admiral	1,317	3.3

Table A.31: Share of the different occupations (C1) for occupation Military

Occupation C1	Frequency	Share
Noble	3,359	16.8
King	2,100	10.5
Peer	1,573	7.9
Prince	1,148	5.7
Emperor	861	4.3
Dynasty	849	4.2
Lord	836	4.2
Duke	819	4.1
Princess	796	4.0
Queen	731	3.7

Table A.32: Share of the different occupations (C1) for occupation Nobility

Occupation C1	Frequency	Share
Founder	5,062	21.3
Recipient	3,159	13.3
Beauty	1,455	6.1
Convict	1,019	4.3
Citizen	818	3.4
Philanthropist	748	3.2
Awarded	642	2.7
Criminal	612	2.6
Crime	610	2.6
Murderer	606	2.6

Table A.33: Share of the different occupations (C1) for occupation Other

Occupation C1	Frequency	Share
Politician	69,058	43.0
Elected	6,871	4.3
Representative	6,324	3.9
President	6,299	3.9
Democrat	4,924	3.1
Republican	4,730	2.9
Political	4,590	2.9
Minister	4,544	2.8
Diplomat	4,426	2.8
Conservative	2,933	1.8

Table A.34: Share of the different occupations (C1) for occupation Politics

Occupation C1	Frequency	Share
Bishop	7,321	20.9
Priest	3,686	10.5
Church	3,134	8.9
Clergy	2,190	6.2
Theologian	1,921	5.5
Archbishop	1,354	3.9
Cardinal	1,294	3.7
Rabbi	1,254	3.6
Missionary	1,239	3.5
Jesuit	1,119	3.2

Table A.35: Share of the different occupations (C1) for occupation Religious

Occupation C1	Frequency	Share
Football	155,681	39.7
Cricket	19,037	4.9
Baseball	18,180	4.6
Ice hockey	16,198	4.1
Rugby	15,222	3.9
Basket	12,993	3.3
Sport	7,114	1.8
Racing	6,931	1.8
Athlete	6,697	1.7
Boxer	6,368	1.6

Table A.36: Share of the different occupations (C1) for occupation Sports

Occupation C1	Frequency	Share
Historian	6,541	8.8
Physician	5,500	7.4
Mathematician	5,073	6.8
Scientist	4,170	5.6
Physicist	4,043	5.5
Economist	3,302	4.5
Philosopher	3,005	4.1
Chemist	2,684	3.6
Astro	2,302	3.1
Botanist	2,231	3.0

Table A.37: Share of the different occupations (C1) for occupation Studies

Occupation C1	Frequency	Share
Sailor	2,137	40.5
Engraver	842	16.0
Worker	779	14.8
Potter	248	4.7
Brewer	169	3.2
Slave	162	3.1
Fisher	120	2.3
Jeweller	118	2.2
Metallurgist	91	1.7
Forester	79	1.5

Table A.38: Share of the different occupations (C1) for occupation Worker

A.5 Census data manipulations

Here we present the most significant manipulations that were needed to compile the Census database presented above. For the United Kingdom: the population of the urban agglomeration of Greater London was added to the second dataset used spanning the time period from 1921-1961. All municipal boroughs within the geographic borders of Greater London were then removed to avoid a double count in population. For the Netherlands a) Male and female populations were combined for the years 1795, 1830, 1849, 1859, 1889, 1869, 1899. b) Certain city and province names were modified to facilitate census record merging. For example, capital letters and hyphens were removed in all city names. Also, province names whose cardinal direction was not specified were modified after city locations were verified. For example, in some cases, the province of Holland was replaced by Nordholland. For Belgium: a) The population of cities within the borders of the urban agglomeration of Brussels (e.g. Uccle, Ixelles, etc.) was combined to yield one observation labeled Brussels (total). The urban area of Brussels was also retained and labeled as Brussels (urban area). Last, for Portugal we decided to consider Lisboa and Porto instead of their respective urban areas Grande Lisboa and Grande Porto respectively.

Le LIEPP (Laboratoire interdisciplinaire d'évaluation des politiques publiques) est un laboratoire d'excellence (Labex). Ce projet est distingué par le jury scientifique international désigné par l'Agence nationale de la recherche (ANR). Il est financé dans le cadre des investissements d'avenir.

(ANR-11-LABX-0091, ANR-11-IDEX-0005-02)

www.sciencespo.fr/liepp

Directeur de publication:
Bruno Palier

Sciences Po - LIEPP
27 rue Saint Guillaume
75007 Paris - France
+33(0)1.45.49.83.61
liepp@sciencespo.fr