# Migration and the epidemiological approach: time and self-selection into foreign ancestries matter*

Simone Bertoli[a], Melchior Clerc[a], Jordan Loper[a], and Èric Roca Fernández[a]

[a] *Université Clermont Auvergne, CNRS, IRD, CERDI, F-63000, Clermont-Ferrand*

**Abstract**

Data on individuals of immigrant origin are used in the epidemiological approach in comparative development for understanding the determinants of cultural traits, and the effects of genetic factors. A widespread presumption in the literature is that this approach is exposed to attenuation bias. We discuss three dimensions of unobserved heterogeneity that are typically overlooked and which can confound the estimation and counteract the attenuation bias. Focusing on the United States, a key context in this literature, we highlight the heterogeneity among natives reporting different foreign ancestries with respect to the average time elapsed since ancestral migration, their spatial concentration, and their attachment to their ancestral identity. These dimensions of heterogeneity vary smoothly across space, oftentimes mirroring the general trend of the variables of interest in this literature, creating a threat to identification. We propose proxies that can be controlled for by the researcher as a bias-reducing strategy.

**Keywords:** comparative development; migration; ancestry; culture; identity choice.

**JEL classification codes**: F22; 012; Z10.

*The authors gratefully acknowledge the comments received from Travis A. Baseler, Martin Fernández, Jesús Fernández-Huertas Moraga, David Gomtsyan, Gordon Hanson, Francesca Marchetta and Gianluca Orefice, and from participants to the 16$^{th}$ AFD-World Bank Conference on Migration and Development (Boston University, October 2023), the 13$^{th}$ Conference on Immigration to OECD Countries (Paris, December 2023) and to an internal seminar at CERDI; they also gratefully acknowledge the support received from the *Agence Nationale de la Recherche* of the French government through the program "*Investissements d'avenir*" (ANR-10-LABX-14-01); the usual disclaimers apply.

Èric Roca Fernández, UCA, CNRS, IRD, CERDI, 26 Av. Léon Blum, F63000, Clermont-Ferrand; email: eric.roca_fernandez@uca.fr (corresponding author).

# 1 Introduction

Understanding the deep-rooted determinants of economic development is a fundamental question that has garnered considerable attention among economists. This has sparked a burgeoning literature, revealing how contemporary outcomes such as economic development, inequality, and individual behaviors are influenced by the persistent characteristics of a distant past. Furthermore, this literature has also provided abundant evidence of the persistence of cultural and genetic traits.

Isolating the long-term causal effect of such traits is a vivid challenge due to the likely emergence of numerous credible confounders. To address this, economists have extensively relied on the so-called epidemiological approach (Fernández, 2011). The underlying logic is clear: isolating the causal effect of cultural, historical or genetic trait on modern outcomes necessitates comparing individuals (or groups of individuals) from different backgrounds residing in the same institutional, legal, economic and geographic context. A common strategy in the literature involves comparing individuals (or groups) within in a single country but originating from different countries. This is justified by the expectation that cultural traits are vertically transmitted across generations (Bisin and Verdier, 2001). Consequently, immigrants or natives of immigrant origin are presumed to retain, at least partly (Giavazzi et al., 2019), the cultural traits of their origin while living in the same environment, thus facing identical incentives and constraints. The United States is a natural focal point in this literature, with both its historical and current migration experience providing valuable identifying variability for the researcher.

This approach faces several challenges that are discussed in the literature. The first challenge relates to the definition of a criterion linking each individual to a single origin. Mostly data-constrained, the literature has used four main criteria: country of birth (Antecol, 2000; Luttmer and Singhal, 2011), maternal or paternal country of birth (see, for instance, Fernández, 2007; Giuliano, 2007; Galor and Savitskiy, 2018; Galor et al., 2020; Giuliano and Nunn, 2021; Ek, 2024), self-reported ancestry (e.g., Antecol, 2000; Guiso et al., 2006; Fernández and Fogli, 2006; Algan and Cahuc, 2010; Alesina et al., 2015; Galor and Özak, 2016; Giavazzi et al., 2019; Arbatlı et al., 2020; Giuliano and Nunn, 2021; Ek, 2021; Galor et al., 2023),[1]

[1] The General Social Survey, which is used in several papers (e.g., Galor and Özak, 2016; Giavazzi et al., 2019; Alesina et al., 2015; Arbatlı et al., 2020), allows identifying natives with foreign-born parents or grandparents, but it does *not* report their countries of birth; the survey rather contains a question on foreign ancestry,

and language (see, for instance, Alesina et al., 2003; Desmet et al., 2017; Giuliano and Nunn, 2021).[2] However, as will be discussed in this paper, except for country of birth, none of these criteria precisely define an individual's ancestry or necessarily identify a single origin: parents can be born in different countries, an individual can speak more than one language at home or report multiple ancestries. Furthermore, the various criteria might end up relating an individual to different origins, e.g., an individual born in France might speak Spanish at home, and declare to be of Italian ancestry. The distinctions between origins are relevant for the econometric analysis to the extent that different origins are characterized by different values of the underlying determinants of cultural traits or of genetic factors.

The multifaceted nature of one's own cultural identity, the fact that migrants from all origin countries are likely to be self-selected along some cultural traits (e.g., long-term orientation and attitude towards risk), and the fact that cultural homophily implies that migrants moving to a given destination from different countries are likely to be more similar than the populations at origin are, all strongly suggest that the epidemiological approach is potentially exposed to an attenuation bias. Indeed, the widely cited review of the literature on the epidemiological approach by Fernández (2011) observes that:

> "It should be noted explicitly that the epidemiological approach is biased towards finding that culture does not matter. As mentioned previously, the fact that parents are only one source of cultural transmission among many and that they may have cultural attitudes that differ from the average ones in the country of ancestry implies that one is more likely to rule the cultural proxy insignificant."

While the expectation of a major attenuation bias is solid-grounded and usually discussed in the literature, we describe and provide novel empirical evidence of an additional potential threat that has been, so far, overlooked and may confound identification in analyses relying on self-reported foreign ancestry for individuals born in the United States. More precisely, this paper demonstrates that natives of different ancestries greatly vary in three dimensions: the average time elapsed since ancestral migration, their spatial concentration and distribution

which can be used to identify the ancestral country of birth for second or third-generation immigrants that chose to report a foreign ancestry.

[2] Additional steps are usually necessary to associate each of these four criteria to the variable(s) whose influence is tested in the econometric analysis. For instance, Giuliano and Nunn (2018, 2021) rely on the mapping of languages for each country from the Ethnologue combined with data on the spatial distribution of the population to aggregate data at the language (ethnicity) level for each country.

within the United States, and their propensity to speak the language of their ancestral country. These three usually *un*observed and *un*controlled[3] dimensions of heterogeneity can confound the identifying variability in the epidemiological approach if they are correlated with the origin-specific variables of interest used in the analysis, thus possibly counteracting the tendency to an under-rejection of the null hypothesis of no effect that represents the common expectation in the literature. The correlation can arise from the absence of major spatial discontinuities across ancestral countries in the three dimensions of heterogeneity that we have just described, and which mostly reflect the evolution over time in the distribution across origin countries of past migration flows to the United States, coupled with the fact that typical variables of interest in the literature only vary smoothly across space.

The time elapsed since ancestral migration shapes the evolution of cultural traits. Guiso et al. (2006) argue that differences in the attitudes towards redistribution of natives of different foreign ancestries could reflect differences in the time elapsed since ancestral migration,[4] and Giavazzi et al. (2019) provide empirical evidence of this. Furthermore, the passage of time also shapes economic outcomes (Abramitzky et al., 2014), and it is associated with a greater dispersion of the population of immigrant origin within the United States. This can be due, for instance, to the fact that the importance of migrant networks (see, for instance Patel and Vella, 2013) declines over time, because the relevance of the local labor demand conditions that shaped the initial spatial distribution of the immigrants fades away, or to the fact natives of immigrant descent move away from more expensive locations where their foreign-born ancestors tend to concentrate (Albert and Monras, 2022). Spatial dispersion can also matter due to within-country differences in economic conditions (Glaeser and Gottlieb, 2009),[5] and in cultural values (Bertrand and Kamenica, 2023), thus calling into question the

[3] To the best of our knowledge, only four papers consider the time elapsed since immigration. Algan and Cahuc (2010), Alesina et al. (2015) and Giavazzi et al. (2019) draw on data from the General Social Survey containing information on both self-reported foreign ancestry and on the number of parents and grandparents that were born abroad to differentiate between first to fourth- (or more) generation immigrants. Ek (2021) checks the validity of his results by using the 1970 US Census data, in which he can isolate the second-generation immigrants. However, none of these strategies accounts for the (potentially sizeable) differential time since parental immigration.

[4] "[...] Americans with British, North European or German ancestors derive from earlier immigrants; hence, more generations were raised in the United States and forged by its culture, absorbing the belief that success is mostly determined by individual actions, which makes government intervention highly undesirable" (Guiso et al., 2006, p. 41).

[5] "[T]he within-country differences in income and productivity are also quite striking. The average income

initial motivation of the epidemiological approach.

These two challenges are compounded by the self-reported nature of foreign ancestry. Natives whose ancestors came from, say, Mexico might intentionally choose not to disclose this foreign lineage and may instead identify as Americans or opt for reporting no specific ancestry. Thus, individuals showing linkages with a given origin might be self-selected along the cultural traits or economic outcomes under analysis. For instance, Farley (1991) argues that descendants of European immigrants to the United States may be less inclined to report a foreign ancestry due to factors such as the passage of time since initial migration, evolving marriage patterns, and a reduced perceived significance of ancestral ties. Duncan and Trejo (2011, 2017) and Antman et al. (2016, 2023) provide evidence of the substantial incidence of what they describe as "ethnic attrition", i.e., the propensity of individuals living in the United States who are first to third-generation immigrants from Latin American or Asian countries *not* to report that they are Hispanics or Asians. Ethnic attrition increases with the time elapsed since ancestral migration, it greatly varies across origin countries, and across individuals, e.g., higher-educated individuals with a Mexican ancestry are significantly less likely to identify themselves as Hispanics.

Surprisingly, studies relying on the epidemiological approach almost invariably refer to concerns related to non-random selection into migration (see, for instance, Jaeger et al., 2010; Beck Knudsen, 2022, for empirical evidence on migrants' selection along cultural traits) but do *not* generally refer to the possible threat posed by natives' non-random selection in a given foreign ancestry, or regard it is a less pressing concern. Giavazzi et al. (2019) represent an exception here, as they explicitly recognize the relevance of the time elapsed since migration,[6] and they are able to differentiate between four generations of immigrants, providing evidence of cultural assimilation across generations. Similarly, the literature does not systematically consider the effects produced by differences in the place of residence within the United States.

per capita in 2007 in the San Francisco metropolitan area was above almost $60,000; the comparable figure for Brownsville, Texas, is under $20,000. Per capita gross metropolitan product (GMP) is more than three times higher in New York than in El Paso." (Glaeser and Gottlieb, 2009, p. 963).

[6] "[V]alues and beliefs depend both on the country of origin of a person's ancestors, as well as on her generation [...]. The country of origin is an important determinant of culture as it encodes the history of a people, encompassing past technological, economic, institutional and cultural environments. The generation of a person is important given that the temporal "distance" from the country of ancestry may be associated with a dilution of the original cultural trait through longer exposure to a different set of economic and social opportunities, to different institutions, and cultural influences."

We draw on data from the 1980 to 2000 population censuses (Ruggles et al., 2023) to explore the answers to the questions on ancestry,[7] and to build three origin-specific variables that capture three relevant dimensions of heterogeneity across natives of foreign ancestry, namely the average time elapsed since ancestral migration, the spatial dispersion within the United States, and the share of natives with ancestors from a non-English-speaking country who still speak their ancestral language.[8] Different countries of ancestries greatly differ with respect to (*i*) the share of the population living in the United States with a given foreign ancestry that is born in the ancestral country, e.g., the share of individuals of Canadian or French Canadian ancestry that report being born in Canada, (*ii*) in their spatial dispersion within the United States.[9] More precisely, countries of origin for which migration flows to the United States mostly occurred in a distant past and characterized by a limited ethnic attrition across generations, have a much lower share of individuals born in the ancestral country, and the natives of immigrant descent are more evenly distributed across states in the United States. For instance, just around 3 percent of individuals of German ancestry in the 2000 census are born in Germany, and their spatial distribution starkly differs from the one presented by Abramitzky and Boustan (2017) for the immigrants born in Germany and in Austria in 1920. Conversely, the share of individuals with Latin American ancestries born in the ancestral country is close to or even above 50 percent, according to the data from the 2000 census, with a major concentration in a few states. Furthemore, (*iii*) natives claiming an ancestry in a Latin American country are substantially more likely to speak their ancestral language (mostly Spanish) at home than natives with a European ancestry in a non-English speaking country.

Our paper makes three distinct contributions to the literature: First, it provides a detailed overview of the different variables that are used in the literature to identify the ancestral origin of the natives, emphasizing the advantages and the analytical challenges related to the various options. As the population census in the United States and the ACS leave little

[7] Since 1980, the population census includes a question, for all individuals irrespective of their citizenship status, related to the respondent's "ethnic origin or descent, roots, or heritage"; the same question is also included in the American Community Survey and in the General Social Survey, while none of them include questions on the parental countries of birth.

[8] We consider English-speaking countries those where English is an official language or where it is spoken by at least 20% of the population, according to Mayer and Zignago (2011).

[9] Foreign-born individuals almost invariably report an ancestry that corresponds to their country of birth, so that more recent migration flows from a given origin are associated with a higher value of the variable.

alternative to the use of the question on the self-reported ancestry that was introduced in 1980, we highlight the threats to identification in the epidemiological approach stemming from unobserved heterogeneity across different origins in the timing of past immigration flows. Second, it provides evidence about the major extent of heterogeneities across natives with distinct self-reported foreign ancestries. Third, it proposes how to build origin-specific proxies using census data or large scale surveys such as the ACS that are informative about the time elapsed since ancestral migration. Such variables can then be introduced in studies using data from other surveys, such as the CPS and the General Social Survey, where the smaller sample sizes prevent a reliable computation of these proxies.

Our paper is purposely descriptive for several reasons. First, census data and surveys do not provide both self-reported ancestry and a factual measure of the origin of one's own ancestors, e.g., the countries of birth of the grandparents or great-grandparents.[10] This implies that we do not have information on the set of possible foreign ancestries of each respondent, and we cannot study the individual-level correlates of the choice of the self-reported foreign ancestry. Second, albeit 10 out of the 11 papers using foreign ancestries that we cited above have already been published, the replication files are publicly available for just three of them (Galor and Özak, 2016; Arbatlı et al., 2020; Giuliano and Nunn, 2021). This hinders our ability to systematically examine how unobserved heterogeneity dimensions among natives with self-reported foreign ancestries may impact published results. Related to this, each of these papers presents many variants of the equation of interest,[11] and this would have exposed any evidence that we would have provided to the (legitimate) concern that we had deliberately selected the specifications to report (see Voth, 2021, on this). Third, any evidence would have been contingent on the choice of the additional origin-specific controls included in each paper, which greatly vary across.

The rest of the paper is organized as follows: Section 2 briefly describes the canonical equation that is brought to the data, and the possible threats to identification related to three distinct criteria that can be adopted to identify the origin of each respondent; Section 3 describes the data sources that we employ; Section 4 presents the results from our descriptive

[10] Fulford et al. (2020) provide an aggregate measure of the ancestral countries of origin for the population of each county in the United States; their approach cannot be deployed at the individual-level.

[11] For instance, the main text of Galor et al. (2023) presents 44 different specifications featuring their origin-specific variable of interest, i.e., terrestrial migratory distance from Addis Ababa, and the results from 60 additional specifications are included in the appendix.

empirical analysis, and Section 5 draws the main conclusions.

## 2 The epidemiological approach

Let us consider a set of individuals or groups of individuals indexed by $i$, residing in the location $k$ within a single country $d$, and let $o$ represent the origin of an individual or group, as defined later. Suppose that cross-sectional data are collected at time $T$. The typical regression that is brought to the data in the epidemiological approach can be written as follows, omitting the subscripts for the (only) country $d$ and survey time $T$:

$$y_{iok} = \alpha w_o + \boldsymbol{\beta}'\boldsymbol{x}_o + \boldsymbol{\gamma}'\boldsymbol{x}_i[+\boldsymbol{\phi}'\boldsymbol{x}_{ok} + d_k + \lambda f(t_i)] + \epsilon_{iok} \tag{1}$$

where $y_{iok}$ is the dependent variable, $w_o$ is the origin-specific variable of interest, and $\boldsymbol{x}_o$ and $\boldsymbol{x}_i$ are two vectors of origin-level and individual-level variables, $\boldsymbol{x}_{ok}$ and $d_k$ are respectively a vector of origin-location control variables and dummies for the area of residence, $t_i$ is the time elapsed since ancestral migration for individual $i$, possibly transformed through a nonlinear function $f(t_i)$, and $\epsilon_{iok}$ is the error term. The notation in Eq. (1) is meant to reflect the fact that $f(t_i)$, $\boldsymbol{x}_{ok}$, and $d_k$ are not systematically included.[12] Similarly, the vector $\boldsymbol{x}_o$ do not include variables measured from the population from the origin $o$ residing in the country of destination $d$.

The coefficient of interest in Eq. (1) is $\alpha$, and the identifying assumption is that $E(w_o \times \epsilon_{iok}|\boldsymbol{x}_o, \boldsymbol{x}_i, [\boldsymbol{x}_{ok}, d_k, f(t_i)]) = 0$. Violations of this assumption arise if (*i*) individuals from different origins descend from migrants who moved to the destination $d$ at different points in time prior to $T$, $\lambda$ is different from 0, and the average origin-specific time since ancestral migration is correlated with $w_o$, or if (*ii*) individuals associated with various origin countries are differently self-selected with respect to unobserved determinants of $y_{iok}$, when the intensity of this non-random selection correlated with $w_o$.

The relevance of these two potential threats to identification clearly depends on the criterion that is chosen to connect each individual (or group of individuals) in the sample to her own origin. Let us focus here on three different criteria, which implies also different sam-

[12] Giuliano and Nunn (2021) and Arbatlı et al. (2020) are exceptions to the norm insofar as the authors include area of residence fixed-effects $d_k$, and the former also includes a vector $\boldsymbol{x}_{ok}$; however, individual time since ancestral migration $t_i$ remains unaccounted for.

ple selection criteria: country of birth, parental (paternal or maternal) country of birth, and country of foreign ancestry.

## 2.1 Country of birth

This criterion to determine the origin $o$ restricts the sample to first-generation immigrants. Thus, it reduces the concerns related to point $(i)$ above,[13] but, conversely, it magnifies the concerns related to point $(ii)$ because immigrants from country $o$ are likely to differ from the stayers along several cultural traits. To provide just an example, migrants might be be positively self-selected with respect to their long-term orientation and negatively with respect to their risk aversion. Thus, the expected value of the error term $\epsilon_{iok}$ in Eq. (1) is likely to vary with the origin $o$.

## 2.2 Parental country of birth

This criterion for determining the origin $o$ restricts the sample to natives with at least one foreign-born parent, i.e., second-generation immigrants. While this criterion accentuates concerns related to unobserved heterogeneity in $t_i$, the time elapsed since the migration to country $d$ of the parents of individual $i$,[14] it can alleviate concerns related to point $(ii)$. The extent of non-random selection in unobservables may be dimished when transitioning from the first to the second generation of migrants (see Giavazzi et al., 2019, on this), under the plausible assumption of imperfect vertical transmission of cultural traits (Bisin and Verdier, 2001).

## 2.3 Country of foreign ancestry

The utilization of self-reported country of foreign ancestry narrows down the sample used for estimating Eq. (1) to native individuals who declare a foreign country as their ancestral origin. This third criterion exacerbates concerns related to unobserved heterogeneity in the time since ancestral migration $t_i$, as the data lacks information on the number of generations that separate individual $i$ from ancestors who migrated to country $d$.

[13] These concerns can be further mitigated by including the years since migration in the vector $\boldsymbol{x}_i$ in Eq. (1).

[14] For instance, the same survey can feature an 80 years old individual born in the United States whose parents migrated at the age of 30, and a young individual born to parents that just arrived into the country. For this example, the difference in time since parental migration may be close to a century.

While this criterion seemingly alleviates concerns related to poin (*ii*) above, as greater values of $t_i$ can futher attenuate the initial migrants' non-random selection in unobservables,[15] this dilution may *not* apply to individuals reporting foreing ancestry. Indeed, some individuals with ancestors from a specific country (e.g., Mexico) might choose not to report any ancestry or to declare an American ancestry or "general heritage" (Hispanic) ancestry, resulting in their exclusion from the estimation sample.[16][17]

For natives of immigrant origin, the choice of reported ancestry reflects the evaluation of costs and benefits of each possible identity (Akerlof and Kranton, 2000) and parental efforts to transmit identity to children (Bisin and Verdier, 2001).[18] In the early stages of migration, migrants from most origin countries faced negative attitudes in the United States, as the following two examples, drawn from Fasani et al. (2019), on Irish and Chinese immigrants suggest:

> "Just outside the US borders were 'hordes of Wild Irishmen ... the turbulent and disorderly of all the world [who come to the United States in order to] distract our tranquillity.' (Massachusetts Representative Harrison Gray Otis, 1797)."; "Chinese immigrants were 'morally the lost debased people on the face of the earth' (Connecticut Senator Orville Platt, 1882), who 'bring every character of vice ... [and would be] injurious in every sense of the world' (Texas Senator Samuel Bell Maxey, 1882)."

Historically, Italians also faced long-lasting negative attitudes (see, for instance, Fouka et al., 2021), as individuals originating from Latin American countries still experience in the United States (see, for instance, Chavez, 2013). Thus, natives of immigrant origin who choose to report a foreign ancestry, despite natives' potential negative attitudes, likely have a stronger attachment to this (costly) identity than their counterparts who decide not to declare this ancestry. Similarly, their intensity of non-random selection in unobservables is likely to

[15] Indeed, the cultural traits of natives with migrant ancestors are more and more determined by a horizontal transmission through social interactions within country $d$ as $t_i$ increases, and they depend less on the country of ancestral origin $o$.

[16] 10% of the population did not report an ancestry in 1990, a figure that jumped to 19% in 2000 (see Brittingham and de la Cruz, 2004, on this), and another substantial share of the population reported a "general heritage", like African or European, that cannot be connected to a country.

[17] These natives would be associated to a different origin if they reported a different foreign ancestry, if some of their ancestors came from another country.

[18] Antman and Duncan (2023) have recently developed a similar theoretical framework to the analysis of identity choice among Native Americans in the United States.

be stronger than for the individuals reporting an ancestry that natives favorably regard.[19]

Returning to Eq. (1), this suggests that natives with different self-reported foreign ancestries are likely heterogeneous regarding the unobserved determinants of cultural traits $y_{iok}$. Consider, for instance, the importance of tradition as a cultural trait, and compare two foreign ancestries in the United States, German and Mexican. German has been the most common ancestry in the United States since the 1990 census, when it accounted for 23 percent of the population (Brittingham and de la Cruz, 2004), while Mexico represents the main recent country of immigration to the United States, but a relatively small share of self-reported foreign ancestries. Natives choosing to report a Mexican ancestry are likely to be more attached to tradition than natives with German ancestry, as the parents of the former have probably exerted a larger effort to ensure the transmission of their identity, given the predominantly negative attitudes towards Hispanics in the United States. The direction of the ensuing bias clearly depends on the sign of the correlation between $w_o$ and $\epsilon_{iok}$.

# 3 Data sources

## 3.1 Population censuses in the United States

Our main data source is the 5 percent sample of the 2000 population censuses in the United States (Ruggles et al., 2023), augmented with the 5 percent samples of the 1980 and 1990 census in selected parts of the analysis.[20,21] The censuses provide a wealth of individual characteristics about all individuals residing in the United States, including undocumented immigrants, and, together with the American Community Survey, it represents a relevant data source in the literature (see, for instance, Fernández and Fogli, 2006; Giuliano and Nunn, 2021; Ek, 2021; Galor et al., 2020, 2023). The United States has historically attracted, and it continues to attract, large numbers of immigrants from various origins, making it an ideal

[19] Duncan and Trejo (2011) and Duncan and Trejo (2017) provide evidence of the extent to which ethnic attrition, defined as the incidence of individuals of a given origin *not* claiming the corresponding identity, varies across origin countries.

[20] Other data sources commonly used in the literature focusing on the United States —the March Supplement of the Current Population Survey and the General Social Survey— allow identifying the maternal and the paternal country of birth, but have a much more limited sample size. This prevents us from building reliable measures of the origin-specific variables that we use in our analysis.

[21] The American Community Survey includes the same variables as the population census.

setting on which to apply the epidemiological approach both to understand the deeply-rooted causes of personal traits, and to investigate how the latter persist. As mentioned before, this approach requires information on the origin of each individual, and the United States census provides four variables informative of that: place of birth, ancestry, spoken language at home (other than English) and parental country of birth.

In the remainder of this paper, we will only exploit variables about place of birth and ancestry to identify an individual's origin. Among the these, only birthplace is universally available. Parental country of birth is available only for individuals co-residing with their parents, implying that the sample is limited in size and self-selected. The information about the language spoken at home is similarly not well-suited to identify one's own origin for two main reasons: first, the language spoken at home is clearly endogenous with respect to migration;[22] second, spoken languages do not, in general, uniquely identify a foreign country, e.g., someone speaking Spanish at home could originate from Mexico or El Salvador.

## 3.2 Variables used in the analysis

We use four main variables in the analysis, namely the birthplace of a respondent, self-reported ancestries, the maternal or paternal birthplace, and the languages other than English spoken at home by a respondent. As discussed above, the first two will be used as sample selection criteria, and to define the country of origin of each individual, while the other two variables will be used in the analysis, but not in the definition of the country of origin. We also describe how we established crosswalks between birthplaces, ancestries and languages.

### 3.2.1 Birthplace

The variable `bpl` describes either the state of birth in the United States, or, in general, the country of birth for foreign-born individuals. This variable is presented in two variants: the first containing country names, while the second, more comprehensive option, includes sub-regional identifiers where applicable. For instance, the latter version includes unique codes for individuals born in specific regions like Aruba or Madeira for the year 2000.

[22] Information about the mother tongue was recorded in the census only until 1970.

### 3.2.2 Maternal or paternal birthplace

This variable can be built combining the information on the variable bpl with the variables momloc and poploc, which provide the identifier (if any) of the co-resident mother and father of each respondent.

### 3.2.3 Self-reported ancestries

The Census Bureau first introduced the ancestry question in the 1980 census, when the questions on the maternal and paternal birthplaces, and on the mother tongue, were abandoned. This question offers respondents the opportunity to specify the "ancestry group with which [she] identifies."[23] Participation in this question is voluntary, and respondents can decide not report any ancestry. The text of the questionnaire includes examples of ancestries, which in 2000 were the following: "Italian, Jamaican, African Am., Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on".[24]

Respondents can report multiple ancestries; more precisely, up to two ancestries in the 1990 and in the 2000 census (and in the ACS). The information is recorded in the variables ancestr1 and ancestr2. The Census Bureau does not provide to the respondents any explicit criterion to determine the order of the ancestries that are reported. Things were different in 1980, when up to three answers were recorded. When this occurred, the ancestries were coded in alphabetical order in the variable ancestr1 only, while the variable ancestr2 was missing.[25] Thus, someone saying she was (in this order) of German and English descent, had German coded for ancestr1, and English for ancestr2. Conversely, someone saying she was (in this order) of German, English and French descent, would have English-French-German coded for ancestr1, and for ancestr2 would be missing. Thus, the third ancestry would alter the order of the first two.

The answers that are coded can encompass both native ancestries, e.g., American, Afro-American or American Indian, and foreign ancestries. Foreign ancestries can relate directly to a country, e.g., Italian, to a regional entity within a country, e.g., Sicilian, or to a supranational entity, e.g., European or Hispanic.[26] The Census Bureau treats in a specific way

[23] See https://usa.ipums.org/usa-action/variables/ANCESTR1#questionnaire_text_section.

[24] The list of examples has evolved over time, as discussed below.

[25] See https://usa.ipums.org/usa-action/variables/ANCESTR1#comparability_section.

[26] The Census Bureau refers to this latter type of ancestries as "general heritages".

the American ancestry: this is considered as a valid ancestry only when no other ancestry is mentioned, and it is discarded otherwise. Thus, an individual saying, for instance, that she is of American and German ancestry would have `ancestr1` reporting German, so that one should interpret the first ancestry as actually being the first *valid* reported ancestry.

The following long quote from Farley (1991) nicely describes a few key features of the question about ancestry:

> "The ancestry question is more complicated in that it depends to a large degree on factual knowledge about the history of one's family, but also requires many people to make a decision about identification: Which one or two of several possible ethnicities does a person report? Undoubtedly, some individuals identify very strongly with a particular ancestry, and will do so regardless of clues on the census form or contemporary political events. Many others, however, may not identify strongly, so their answers may depend on ephemeral events." (Farley, 1991, p. 414).

With respect to ephemeral events, these can also be related to the census questionnaire itself. Farley (1991) observes that the ancestry question followed a question on English proficiency in 1980, which might have induced respondents without a strong attachment to an ancestry or with multiple ancestries to report an English ancestry. Similarly, the list of ancestries provided as examples has changed over time, and the changes introduced between 1980 and 1990 have exerted a substantial influence on the distribution of ancestries Rosenwaike (1993). Thus, one should keep in mind that the question about ancestries does not, differently from the question on the birthplace, elicit factual information, but rather a subjective judgement, that is closely intertwined with a choice concerning one's own identity.

### 3.2.4 Language spoken at home

The variable `language` codes the language other than English spoken at home by each respondent aged 5 and above. Individuals who only speak English at home are coded as `English`. This variable provides a limited coverage of indigenous languages of foreign countries.

We establish a crosswalk between the variables `bpl`, `ancestr1`, `ancestr2`, and `language` . More precisely, we associate to each reported ancestry the codes corresponding to birthplace, e.g., we associate Belgian and Flemish ancestries to the code of the variable `bpl` corresponding to Belgium. Then, we associate to each ancestry any language that is either an official language of the associated country or that is spoken by at least 20 percent of its population, drawing on the data from Mayer and Zignago (2011). This allows us, for instance, identifying the respondents of a given foreign ancestry that are born in the ancestral country, or that speak a language associated to the ancestral country at home.

# 4 Empirical analysis

We describe here the various steps of our empirical analysis.

## 4.1 Identifying individuals' origin

Following the standard practice in the literature, we identify one's own origin on the basis of the first reported ancestry only.[27] As our objective is to associate the answers to a (single) foreign country, we do not consider "general heritage" ancestries that cannot be reliably linked to a country: American, Afro-American, American Indian, Eskimo, European, or Latin American, among others. Then, we aggregate distinct ancestries at the level of a country. This requires first aggregating more detailed ancestries, e.g., associate English, Northern Irish, Scottish, Scotch Irish and Welsh ancestries to Great Britain, or Acadian, Canadian and French Canadian to Canada.

We begin our analysis with an exploration of ancestry for the 1980–2000 population censuses. Overall, in the 2000 census, around 58.6 percent of the respondents report at least one foreign ancestry that we can exploit in the analysis. Focusing on natives only, this figure drops slightly to 55.7 percent. We can match ancestries to 109 different foreign countries. Almost all European and American countries are separately identified. At the same time, this is not the case, with just a few exceptions, for Africa.[28]

[27] A possible alternative is, for respondents whose first ancestry is either a native ancestry or a general heritage, would be to use the second reported ancestry, when the latter is non-missing and it can be connected to a foreign country; this alternative is immaterial, as most individuals with a first ancestry that we cannot exploit do not have a second ancestry that can be connected to a country.

[28] Brittingham and de la Cruz (2004) observes that the share of individuals reporting an African general heritage increased from 246,000 to 1.2 million between 1990 and 2000.

## 4.2 Main foreign ancestries among natives in 2000

If we focus on the 2000 census and on the natives with a foreign ancestry that we can associate with a foreign country, the five main countries of first ancestry are Germany (12.3 percent considering all natives, 22.1 percent among people with at least one reported ancestry), Great Britain (9.7 and 17.6 percent), Ireland (7.6 and 13.6 percent), Italy (4.7 and 8.5 percent), and Mexico (3.9 and 7.0 percent), representing in total 38.4 percent of the natives (69.0 if we consider only individuals reporting at least one ancestry). Notice that just seven more countries represent at least 1.0 percent of the ancestries of the native population. Figure 1 plots, on a world map, the share of natives reporting a given country as their first ancestry, while Figure 2 repeats the same exercise also using the information on the second ancestry.[29,30]

Figure 5 reports the share of natives with a given first ancestry who also report a second ancestry, and Figure 6 reports the share of natives with a given ancestry (first or second) who also report another ancestry. On average, countries with a more distant migration history to the United States have a higher incidence of multiple ancestries, which capture the occurrence of mixed marriages, i.e., marriages between individuals of different ancestries, in the previous generations.

Figure 7 associates to each Public Use Microdata Area (PUMA) the most frequent ancestry among natives,[31] revealing that 23 distinct ancestries represent the main ancestry in at least one PUMA. The geographic distribution of the most frequently reported ancestries among native-born Americans at the PUMA level matches what we would expect based on historical settlement patterns and migration trends. For example, Mexican ancestry is dominant in PUMAs located near the US-Mexico border in states like Texas, California, Arizona and New Mexico. This aligns with the proximity to Mexico and history of Mexican immigration and settlement in these regions. Similarly, French ancestry is concentrated in PUMAs across Louisiana, consistent with France's colonization of that region in the 18th century, while Norwegian ancestry is mostly situated in the northern parts of North Dakota and Minnesota, reflecting the large waves of Norwegian immigrants settling in the Midwest in the late 19th century, especially in North and South Dakota, possibly because they searched for locations

[29] Notice that this implies that, except for the American ancestry, ancestries are no longer mutually exclusive, and the sum of the shares across all countries now exceeds 100 percent.

[30] As Figures 1 and 2 do not allow visualizing differences in the share of the various ancestries when these are low, 3 and 4 provide the same information using a logarithmic scale.

[31] We associate American, Native American and Afro-American ancestries to the United States.

with similar climatic conditions with respect to Norway (Obolensky et al., 2024).

PUMA-level ancestry data underscores how the time elapsed since migration can vary greatly across reported ancestries, in ways that an epidemiological framework must account for. The fact that German was the most commonly reported ancestry in 1990, with roots tracing back over a century for many of those populations, provides a clear example. Unlike more recent immigrant ancestries, descendants of 19th century German migrants have had longer to geographically disperse across with the United States across generations. We see evidence of this increased dispersion in the PUMA data, with high shares of German ancestry appearing more widely distributed across different regions, as compared to more concentrated patterns of ancestries with closer ancestral migration times, like Mexican nearer the southern border.

## 4.3 Self-reported ancestry varies over time

Brittingham and de la Cruz (2004) documents major changes in the share of the population that identifies with different foreign ancestries between the 1990 and the 2000 census. Notably, the share of the population reporting a German ancestry (the most common ancestry) declined from 23.3 to 15.2 percent of the population, with an absolute decline from 57.9 to 42.8 million individuals.[32] This reduction might reflect demographic events, and notably the death of older cohorts of the population who reported a German ancestry in 1990, coupled with limited incoming migration flows from Germany, and a lower propensity of new cohorts to report a foreign ancestry. Our analysis of the data reveals that the propensity to report a German ancestry greatly varies across censuses for natives born in a given year. More precisely, Figure 8 plots the share of natives born between 1940 and 1979 reporting German as a first ancestry in the 1980, 1990 and 2000 census. The restrictions on the year of birth are meant to minimize the influence of demographic events, as no individual in our analysis is aged above 60 in 2000. For each cohort, the share with a German ancestry markedly increases from 1980 to 1990, and then it abruptly declines in 2000. The changes in the treatment of multiple ancestries from 1980 to 1990 (see the discussion on this in Section 3.1) might explain the increase, but not the ensuing decline. The evolution over time in the share of natives reporting a German

[32] The difference with the figure that we reported above for Germany based on the 2000 census is due to our focus on the native rather than total population, and to the exclusion of individuals reporting either a native ancestry, or a general heritage as first ancestry.

ancestry might also be related to the fact in the enumerators mentioned German ancestry in the fourth place of the list of examples of ancestries in 1980 and in the first place in 1990 (see Rosenwaike, 1993, for evidence on the influence of this change in wording), while German ancestry was not even mentioned in 2000.

These numbers suggest that one should be cautious when exploiting data from the different survey waves to identify individuals' origins, and show the robustness of results when estimating epidemiological regressions using different survey waves separately.

While we have just provided an example related to the most common foreign ancestry, our point is more general. The influence of the ephemeral nature of the self-reported ancestry on the econometric analysis depends on what drives the variation in the answers that a cohort of individuals gives at different points in time. Furthermore, major swings in the incidence of a given foreign ancestry for a given cohort of natives, as we have documented for Germany, strongly suggest a limited attachment to this specific identity.

## 4.4 Capturing unobserved heterogeneity across foreign ancestries

Lacking information on time since ancestral migration for all the origin countries for the individuals who self-report a foreign ancestry, we propose two variables aimed at proxying such unobserved heterogeneity, and a third variable which is a revealed measure of the attachment to one's own ancestry.

The first variable, that we denote as $v_o^1$, is the probability that an individual (either native or foreign-born) claiming a given ancestry was actually born in that ancestral country. This probability will be higher for more recent waves of migration, and for ancestral countries where ethnic attrition is lower. As the generations pass and migration events recede further into the past, fewer ancestry claimants will have been born abroad, even if they still identify culturally with their ancestral heritage. In essence, the variable is a proxy for the number of generations since the initial migration or migrations occurred.

The second measure, which we denote as $v_o^2$, exploits the strong network effects that influence where migrants initially locate within a host country. These network effects lead to clustering, with immigrants initially concentrating in just a few selected areas. Over time, as the generations pass, these network effects diminish as the migrants' descendants integrate more fully into the host society and relocate freely within its borders. To measure the geographic concentration of natives claiming each ancestry within the United States, we calculate a Herfindahl-Hirschmann index for each group. Consistent with the prior discussion, more recent migrant waves should display higher levels of geographical concentration.

The third measure, called $v_o^3$, is represented by the share of natives of a given foreign ancestry reporting speaking at home the language of their ancestral country. This variable, which represents a revealed measure of the importance of tradition according to Giuliano and Nunn (2021), can be meaningfully computed only for natives with an ancestry in a country in which English is not an official language.[33]

### 4.4.1 Share born in the ancestral country

For each foreign ancestry, we compute the share of individuals born in the ancestral country, e.g., the share of individuals of Italian ancestry that are born in Italy, that we denote as $v_o^1$. This variable captures the share of first-generation immigrants among all individuals of a given ancestry. Hence, it is informative of the average time elapsed since ancestral migration even when we restrict the sample to native-born individuals only. The analysis of the data clearly reveals that $v_o^1$ is substantially lower for European than for Latin American countries of origin, and among Latin American countries it is lower for Mexico and other Central American countries, that have a longer history of migration to the United States (see Figure 9).

To provide a few examples, this variable is equal to 2.1 percent for Germany, 0.7 percent for Ireland, 1.2 percent for Great Britain, and 3.5 percent for Italy, while it jumps to 47.2 percent for Mexico and 72.7 percent for El Salvador. The stark differences between European and Latin American countries are consistent with the different timing of migration to the United States from these regions (see notably Figure 2 in Abramitzky and Boustan, 2017), but this proxy is also informative of the differences in timing within regions, as Mexican migration preceded migration from other Latin American countries. Notice that $v_o^1$ is, strictly speaking and by construction, uninformative about the average time since ancestral migration among the natives of a given ancestry. However, when examining data on children co-residing with their parents, it becomes apparent that $v_o^1$ can serve as a reliable proxy, at least for the subsequent generation. This is evident in Figure 10, where we observe that children who identify with ancestral roots spanning a long history of migration usually have parents who were *not* born in the ancestral country. On the other hand, children who identify with more recent migratory waves typically have at least one parent who was born in the corresponding

[33] The accuracy of this variable clearly depends on the coverage of ancestral languages in the data.

foreign country. The correlation between this variable and $v_o^1$ is 0.879 Thus, because different ancestries display a differentiated (average) time since migration, the latter could confound the estimates when relying on the epidemiological approach. To illustrate, suppose one aims to comprehend English proficiency and establish connections with underlying attributes of the ancestral homeland. In this scenario, the estimates would be compromised since earlier waves of migration would have had more opportunities to assimilate and acquire English proficiency.

### 4.4.2 Spatial concentration of natives in the United States

We use the shares $s_{ok}$ to compute an origin-specific Herfindahl-Hirschman index of spatial concentration, that we denote as $v_o^2$. This index corresponds to the probability that two randomly drawn natives with the same ancestry reside in the same area. We compute it at the State rather than the PUMA level, to ensure a sufficient number of observations per area from each origin. The values of the index range between 3.5 percent (Great Britain) and 72.5 percent (Anguilla),[34] with an average and median value standing at 18.2 and 13.3 percent respectively. The values of $v_o^2$ is much lower for European countries than for Latin American countries, as can be seen from Figure 11. For instance, we have that $v_{MEX}^2 = 28.3$ percent, and the Spanish-speaking Latin American country with the lowest concentration is Bolivia, with $v_{BOL}^2 = 11.4$ percent. In contrast, a country with older migration waves like Germany displays a value $v_{DEU}^2 = 4.8$ percent. Again, the large disparities across origins speak to a differential time since migration, with higher concentration persisting for groups arriving more recently due to stronger network effects that initially influence location choice within host countries.

### 4.4.3 Share of natives speaking the ancestral language

The choice to speak the ancestral language at home can be regarded, following Giuliano and Nunn (2021), as a revealed measure of one's own attachment to the identity associated to a foreign ancestry. The construction of this measure, that we denote as $v_o^3$, is clearly meaningful only for ancestries that do not correspond to countries where English is an official language, or where at least 20 percent of the population speaks English (Mayer and Zignago, 2011) and so this can be defined only for 69 out of the 109 foreign ancestries. Figure 14 plots the value of $v_o^3$ for all foreign (first) ancestries corresponding to a non-English speaking country. This

[34] Notice that the high percentage for Anguilla is mechanical, as very few natives, i.e., 95 individuals in the five percent sample from the 2000 Census, report an ancestry in this tiny Caribbean island.

world map clearly reveals that no more than 10 percent of the natives of European ancestries (except for Spanish ancestry) speak their ancestral language at home, while the corresponding share of Latin American ancestries is close or above 50 percent. Figure 15 considers both first and second ancestries in the definition of $v_o^3$. Again, these findings highlight the importance of accounting for differential opportunities for assimilation across ancestry groups over time. Estimates that pool together groups with varied migration histories risk being compromised, as earlier migrant waves would have had more time to acquire English proficiency and adapt to mainstream American culture. Alternatively, and to keep with the narrative in Giuliano and Nunn (2021) where climatic variability is the underlying variable determining attachment to an ancestry, the most changing weather is found in rich countries. As a consequence, migrants from these areas would have had more time living in English-dominant environments to assimilate linguistically through generational language transmission and shift.

### 4.4.4 Correlations and patterns of spatial variation

We explore here the correlation between our three origin-specific variables that we built from the 2000 census, notably the share of the population (natives and immigrants) with an ancestry in the foreign country $o$ born in the ancestral country ($v_o^1$), the Herfindahl-Hirschmann Index of spatial concentration of the native population with an ancestry in the foreign country $o$ ($v_o^2$), and the share of the native population with an ancestry in the foreign country $o$ speaking the ancestral language at home ($v_o^3$).

The correlation between $v_o^1$ and $v_o^2$ can be computed for all the 109 foreign ancestries that we use in the analysis, while the correlations with $v_o^3$ are restricted to the 69 countries of foreign-ancestry where English is not an official language. The correlations are computed by weighting observations by the number of natives with an ancestry in each country $o$. This is the relevant measure, whenever the econometric analyses in the epidemiological approach relying on a variant of Eq. (1) are conducted with individual-level data. Furthermore, weighting observations is important, to avoid giving too much influence to ancestries reported by a limited number of natives, which mechanically have a high value of $v_o^2$.[35]

The correlation between $v_o^1$ and $v_o^2$ stands at 0.863, the one between $v_o^1$ and $v_o^3$ at 0.402, and the one between $v_o^2$ and $v_o^3$ at 0.500. These high correlations provide additional credence

[35] However, we obtain similar results when computing unweighted correlations, which are relevant for analysis conducted at the ancestry-level.

to the proxies we built to capture unobserved heterogeneities of individuals of foreign origin, that should be controlled for in the epidemiological approach.

At least eight out of the ten origin countries with the lowest values of by $v_o^1$, $v_o^2$ or $v_o^3$ are European countries, whose migrants typically moved to the United States in a distant past, while Latin American countries tend to be among the ones with the highest values of these three variables. Indeed, this confirms that foreign countries with a more distant history of migration to the United States are, on average, characterized by lower values of $v_o^1$, $v_o^2$, and $v_o^3$, i.e., these ancestries are mostly composed by natives, who are more spatially dispersed across states, and which have a lower tendency to speak their ancestral language at home.

A simple (weighted) regression of $v_o^1$ on dummies for the five continents produces an $R^2$ of 0.804, while the corresponding $R^2$ when the dependent variable is $v_o^2$ stands at 0.700. Thus, most of the variability in these two measures is across rather than within continent. However, this is not the case with $v_o^3$, as the $R^2$ stands at 0.101, i.e., this variable is characterized by a greater variability within rather than across continents. This, in turn, implies that the inclusion of continental dummies in Eq. (1) would absorb a substantial portion of the possible confounding effect of differences across origin countries in $t_i$, but not in the attachment to the foreign ancestry, as captured by the choice of the language spoken at home.

### 4.4.5 Individual-level regressions

To better understand the relationship between $v_o^1$, $v_o^2$ and $v_o^3$, we perform a simple exercise inspired by Oster (2017) using individual-level data from the 5 percent sample of the 2000 census for natives with a first foreign ancestry that we can match to a country, and whose first and second ancestry do not correspond to one of the 40 countries (out of 109) where English is either an official language or a language spoken by more than 20 percent of the population according to Mayer and Zignago (2011). For 1,204,114 natives aged 25 to 65, residing in a metropolitan area and whose first ancestry corresponds to 69 distinct foreign countries, we define a dummy variable taking the value of 1 if a native speaks at home a language associated to her ancestral country, and 0 otherwise. This dummy variable represents a plausible measure of the attachment to the cultural identity of the ancestral country that is reported, and its average value stands at 10 percent. Then, we regress this variable on a set of individual-level variables $\boldsymbol{x}_i$ and on dummies $d_k$ for the metropolitan area of residence of the native $i$, and we compute the $R^2$ of this regression, which stands at 0.219 (see Table 1). Thus, individual-level variables and dummies for the area of residence explain 21.9 percent of the variability in the dependent variable. Then, we augment this specification with 69 dummies $d_o$, one for each ancestral country. This increases the $R^2$ to 0.466 and represents the maximum explanatory power of that any origin-specific variable can exhibit. We then remove the dummies $d_o$, and we add to the regression either $v_o^1$ (Column 1) or $v_o^2$ (Column 2), or both, without or with dummies for the continent of the ancestral country $o$ (Columns 3 and 4, respectively). By itself, $v_o^1$ explains $(0.405-0.219)/(0.466-0.219) = 75.3$ percent of the variability across countries in this measure of cultural attachment. In other words, the average time elapsed since ancestral migration, which is associated with a lower value of $v_o^1$, (strongly) negatively correlates with the probability of speaking at home a language associated to the ancestral country. Table 1 reveals that a 1 percentage point increase in $v_o^1$ is associated to a 1.012 percentage points increase in the probability of speaking an ancestral language at home. Similarly, a greater spatial concentration ($v_o^2$) is associated with a greater probability of speaking the ancestral language, and this variable alone accounts for $(0.381 - 0.219)/(0.466 - 0.219) = 65.6$ percent of the between-ancestries variability. These results suggest that the two dimensions of unobserved heterogeneity are closely intertwined, with a greater average time elapsed since ancestral migration being, for those that still report a foreign ancestry, associated with a lower intensity of self-selection into such an identification.

Table 1 also reveals that five continent dummies suffice to explain $(0.432-0.219)/(0.466-0.219) = 86.2$ percent of the variability across foreign ancestries in the propensity to speak a language associated to the ancestral country. This is consistent with the fact that, on average, the ancestors of native of European ancestry arrived to the United States long before the ancestors of the natives of Latin American ancestry. This advocates for the inclusion of continent fixed-effect as a standard practice when using the epidemiological approach, as it represents an easy-to-implement second-best to control for unobserved heterogeneities.[36]

[36] However, out of 11 papers using foreign ancestries that we cited above, only 3 use continent or region of origin fixed effects in at least one of their specifications (Arbatlı et al., 2020; Giuliano and Nunn, 2021; Galor et al., 2023)

Table 1: Additional information contained in the time elapsed since ancestral migration

| | Speaking anc. lang. | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Share born in anc. country ($v_o^1$) | 1.012$^{***}$ | | 0.749$^{***}$ | 0.277$^{***}$ |
| | (0.003) | | (0.005) | (0.007) |
| Herfindahl index ($v_o^2$) | | 1.955$^{***}$ | 0.642$^{***}$ | 0.274$^{***}$ |
| | | (0.006) | (0.010) | (0.008) |
| $R^2$ | 0.405 | 0.381 | 0.410 | 0.435 |
| Max. $R^2$ | 0.466 | 0.466 | 0.466 | 0.466 |
| Continent $R^2$ | 0.432 | 0.432 | 0.432 | 0.432 |
| Ind. controls $R^2$ | 0.219 | 0.219 | 0.219 | 0.219 |
| Individual controls | Yes | Yes | Yes | Yes |
| Continent fixed effect | No | No | No | Yes |
| Mean of dep. var. | 0.10 | 0.10 | 0.10 | 0.10 |
| Number of ancestral countries | 69 | 69 | 69 | 69 |
| Observations | 1,204,114 | 1,204,114 | 1,204,114 | 1,204,114 |

*Notes:* This Table reports the results of estimating Eq. 1 on individual-level data from the 2000 US Census. The dependant variable is a dummy variable that equals one if the language spoken at home is either an official language or a language spoken by more than 20 percent of the population (according to Mayer and Zignago 2011) in the country associated with the first ancestry of the respondent, or the country associated with the second ancestry of the respondent, if available. The individual controls include: categorical variables for age, sex, education, marital status and rural/urban status, and a dummy for the area of residence. Column 1) introduces the share of individuals born in the respondent's ancestral country as main control ($v_o^1$), Column 2) employs the Herfindahl-Hirschmann concentration index for respondents' reported ancestry, computed at the State level ($v_o^2$), Column 3) simultaneously introduces $v_o^1$ and $v_o^2$ while Column 4) further adds continental fixed effects. $R^2$ denotes the $R^2$ of each regression, Max. $R^2$ reports the maximum $R^2$ that regressions can feature when a full set of ancestry controls are included, Continent $R^2$ reports the $R^2$ of regressions including continent fixed effects *instead of* variables $v_o^1$ and $v_o^2$ to assess the explanatory power of continental dummies and contrast it to that of our main variables of interest. Finally, Ind. controls $R^2$ indicates the $R^2$ of regressions featuring only individual-level controls. $^{*}$, $^{**}$ and $^{***}$ denote significance at the 10, 5 and 1 percent level, respectively.

# 5 Concluding remarks

What does the descriptive evidence that we have presented in the previous section imply for the epidemiological approach? A large (and growing) set of influential papers have relied on the first foreign ancestry of natives in the United States to identify their country of origin. This choice reflects, to a large extent, binding data constraints, as the population census (since 1980) and the ACS offer no alternative to the use of this variable. However, this variable does not reflect factual information, but it is rather based on a subjective decision concerning one's own identity. In this light, the literature points out an attenuation bias in the epidemiological approach: relying on self-reported foreign ancestry for US-born individuals to construct origin-country variables would tend to stifle identifying variability due to selective migration.

This paper paints a more nuanced picture, unraveling additional biases usually overlooked in literature using the epidemiological approach. It reveals that natives with different foreign ancestries sharply differ along three intertwined dimensions: the time since ancestral migration, spatial concentration, and a revealed measure of attachment to their ancestral identity. The first two dimensions are likely to determine the relative importance of the vertical and of horizontal interactions in the transmission of cultural traits, while the third dimension can shape the strength of vertical transmission, which the epidemiological approach focuses on. A longer time since ancestral migration, and a greater spatial dispersion in the United States are both likely to reduce the importance of vertical transmission, which is also weakened when a native has a limited attachment to the reported foreign ancestry. Our paper demonstrates that German migration to the US was, on average, older than Mexican migration, and individuals of German ancestry were more geographically dispersed than their Mexican counterparts within the US. Also, one can reasonably assume that individuals of German ancestry face different incentives to maintain language (i.e. a key aspect of their ancestral identity) than individuals of Mexican ancestry. With German being almost useless in the United States and Spanish being widespread (natives of Spanish origin are the Europeans most likely to speak their ancestral language at home), the former have a lower propensity to maintain their language, indicating a lower attachment to their origin identity.

In sum, this paper challenges the widespread presumption in the comparative economic development literature adopting the epidemiological approach that any significant association between a cultural trait and an underlying determinant, or between genetic factors and economic outcomes represents a strong empirical evidence, as this approach would be exposed to a substantial attenuation bias due to selective migration. Individuals of foreign ancestry differ along at least three additional dimensions usually overlooked, which may confound epidemiological identification if they correlate with the origin-specific variable of interest (e.g., cultural trait) and the outcome of interest —a genuine concern. Including the proxies built in this paper as additional controls, along with continental dummies for the ancestral countries, offers a straightforward and effective way to check whether the analysis is sensitive to the unobserved heterogeneities we have brought to light.

# References

Abramitzky, R. and L. Boustan (2017): "Immigration in American Economic History," *Journal of Economic Literature*, 55, 1311–45.

Abramitzky, R., L. P. Boustan, and K. Eriksson (2014): "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 122, 467–506.

Akerlof, G. A. and R. E. Kranton (2000): "Economics and identity," *Quarterly Journal of Economics*, 115, 715–753.

Albert, C. and J. Monras (2022): "Immigration and Spatial Equilibrium: The Role of Expenditures in the Country of Origin," *American Economic Review*, 112, 3763–3802.

Alesina, A., Y. Algan, P. Cahuc, and P. Giuliano (2015): "Family Values and the Regulation of Labor," *Journal of the European Economic Association*, 13, 599–630.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8, 155–194.

Algan, Y. and P. Cahuc (2010): "Inherited trust and growth," *American Economic Review*, 100, 2060–2092.

Antecol, H. (2000): "An examination of cross-country differences in the gender gap in labor force participation rates," *Labour Economics*, 7, 409–426.

Antman, F., B. Duncan, and S. J. Trejo (2016): "Ethnic attrition and the observed health of later-generation Mexican Americans," *American Economic Review*, 106, 467–471.

Antman, F. M. and B. Duncan (2023): "American Indian Casinos and Native American Self-Identification," *Journal of the European Economic Association*, 21, 2547–2585.

Antman, F. M., B. Duncan, and S. J. Trejo (2023): "Hispanic Americans in the Labor Market: Patterns over Time and across Generations," *Journal of Economic Perspectives*, 37, 169–198.

Arbatli, C. E., Q. H. Ashraf, O. Galor, and M. Klemp (2020): "Diversity and conflict," *Econometrica*, 88, 727–797.

Beck Knudsen, A. S. (2022): "Those Who Stayed: Selection and Cultural Change in the Age of Mass Migration," Mimeo, https://annesofiebeckknudsen.com/wp-content/uploads/2022/08/thosewhostayed.pdf.

Bertrand, M. and E. Kamenica (2023): "Coming apart? Cultural distances in the United States over time," Tech. Rep. 4.

Bisin, A. and T. Verdier (2001): "The economics of cultural transmission and the dynamics of preferences," *Journal of Economic Theory*, 97, 298–319.

Brittingham, A. and G. P. de la Cruz (2004): "Ancestry: 2000," US Census Bureau, 2000 Census Brief, available at https://www.census.gov/history/pdf/ancestry.pdf.

Chavez, L. R. (2013): *The Latino Threat: Constructing Immigrants, Citizens and the Nation*, Stanford University Press, second edition.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2017): "Culture, Ethnicity, and Diversity," *American Economic Review*, 107, 2479–2513.

Duncan, B. and S. J. Trejo (2011): "Tracking intergenerational progress for immigrant groups: The problem of ethnic attrition," *American Economic Review*, 101, 603–608.

——— (2017): "The complexity of immigrant generations: Implications for assessing the socioeconomic integration of Hispanics and Asians," *ILR Review*, 70, 1146–1175.

Ek, A. (2021): "Cross-country differences in preferences for leisure," *Labour Economics*, 72, 102054.

——— (2024): "Cultural Values and Productivity," *Journal of Political Economy*, 124, 295–335.

Farley, R. (1991): "The New Census Question about Ancestry: What Did It Tell Us?" *Demography*, 28, 411–429.

Fasani, F., G. Mastrobuoni, E. G. Owens, and P. Pinotti (2019): *Does immigration increase crime*, Cambridge University Press.

Fernández, R. (2011): "Does culture matter?" in *Handbook of Social Economics*, ed. by J. Benhabib, A. Bisin, and M. O. Jackson, Elsevier, vol. 1, 481–510.

Fernández, R. (2007): "Women, Work, and Culture," *Journal of the European Economic Association*, 5, 305–332.

Fernández, R. and A. Fogli (2006): "Fertility: The Role of Culture and Family Experience," *Journal of the European Economic Association*, 4, 552–561.

Fouka, V., S. Mazumder, and M. Tabellini (2021): "From Immigrants to Americans: Race and Assimilation during the Great Migration," *Review of Economic Studies*, 89, 811–842.

Fulford, S. L., I. Petkov, and F. Schiantarelli (2020): "Does it matter where you came from? Ancestry composition and economic performance of US counties, 1850–2010," *Journal of Economic Growth*, 25, 341–380.

Galor, O., M. Klemp, and D. Wainstock (2023): "Roots of Inequality," NBER Working Paper 31580.

Galor, O. and O. Özak (2016): "The Agricultural Origins of Time Preference," *American Economic Review*, 106, 3064–3103.

Galor, O. and V. Savitskiy (2018): "Climatic Roots of Loss Aversion," NBER Working Paper 25273.

Galor, O., O. Özak, and A. Sarid (2020): "Linguistic Traits and Human Capital Formation," *AEA Papers and Proceedings*, 110, 309–13.

Giavazzi, F., I. Petkov, and F. Schiantarelli (2019): "Culture: Persistence and evolution," *Journal of Economic Growth*, 24, 117–154.

Giuliano, P. (2007): "Living arrangements in Western Europe: Does cultural origin matter?" *Journal of the European Economic Association*, 5, 927–952.

Giuliano, P. and N. Nunn (2018): "Ancestral characteristics of modern populations," *Economic History of Developing Regions*, 33, 1–17.

——— (2021): "Understanding cultural persistence and change," *The Review of Economic Studies*, 88, 1541–1581.

Glaeser, E. L. and J. D. Gottlieb (2009): "The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States," *Journal of Economic Literature*, 47, 983–1028.

Guiso, L., P. Sapienza, and L. Zingales (2006): "Does culture affect economic outcomes?" *Journal of Economic Perspectives*, 20, 23–48.

Jaeger, D. A., T. Dohmen, A. Falk, D. Huffman, U. Sunde, and H. Bonin (2010): "Direct evidence on risk attitudes and migration," *The Review of Economics and Statistics*, 92, 684–689.

Luttmer, E. F. P. and M. Singhal (2011): "Culture, context, and the taste for redistribution," *American Economic Journal: Economic Policy*, 3, 157–179.

Mayer, T. and S. Zignago (2011): "Notes on CEPII's distances measures: The GeoDist database," Working Papers 2011-25, CEPII.

Obolensky, M., M. Tabellini, and C. Taylor (2024): "Homeward Bound: How Migrants Seek Out Familiar Climates," Working Paper 32035, National Bureau of Economic Research.

Oster, E. (2017): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

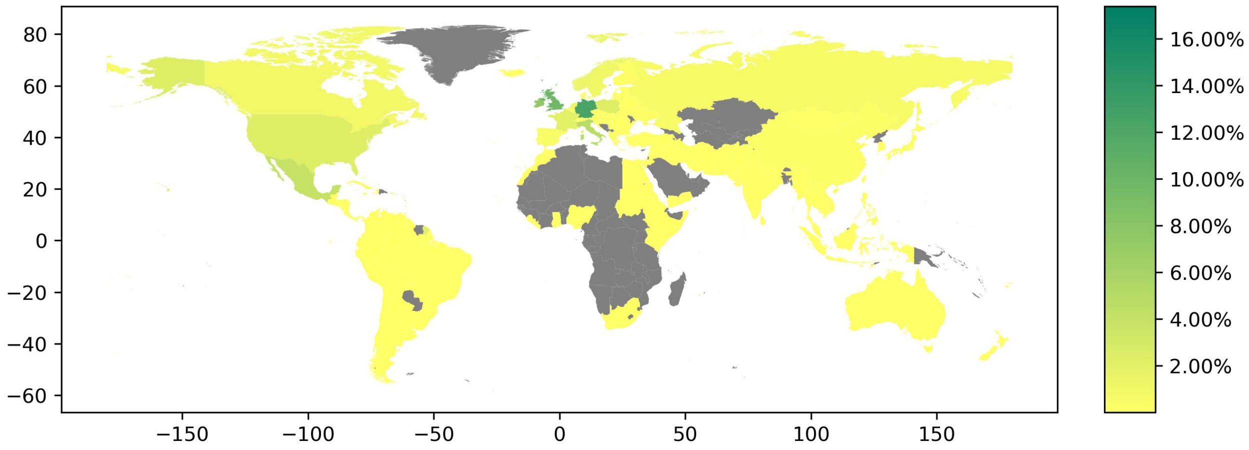
Patel, K. and F. Vella (2013): “Immigrant networks and their implications for occupational choice and wages,” *Review of Economics and Statistics*, 95, 1249–1277.

Rosenwaike, I. (1993): “Ancestry in the United States Census, 1980-1990,” *Social Science Research*, 22, 383–390.

Ruggles, S., S. Flood, M. Sobek, D. Brockman, G. Cooper, S. Richards, and M. Schouweiler (2023): “IPUMS USA: Version 13.0 [dataset],” Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V13.0.

Voth, H.-J. (2021): “Persistence–myth and mystery,” in *The Handbook of Historical Economics*, ed. by A. Bisin and G. Federico, Elsevier, 243–267.

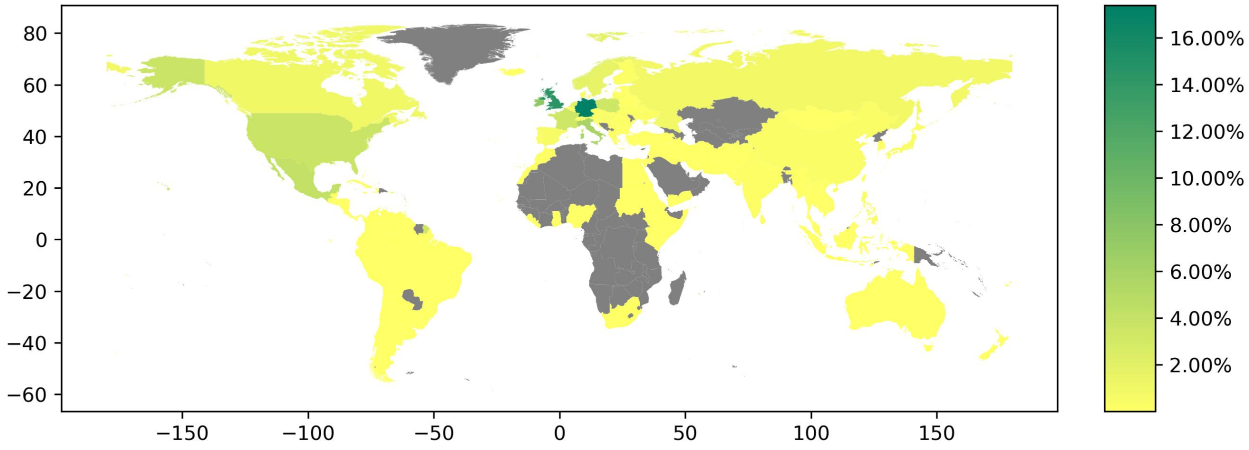
Figure 1: Proportional Distribution of Native Ancestry by Country in 2000



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first ancestry. In the case of the United States, both American and Native American ancestries are attributed to it.

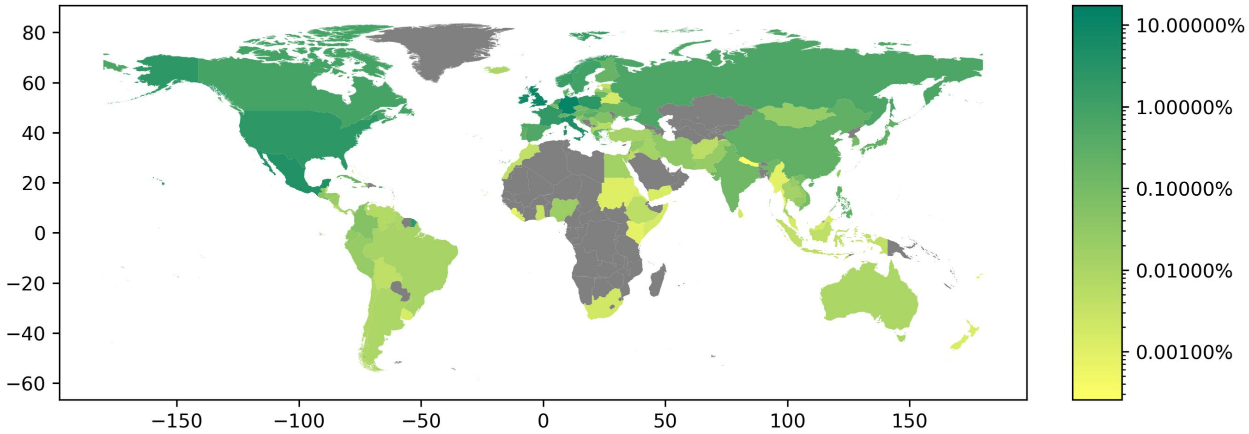
Figure 2: Proportional Distribution of Native Ancestry by Country in 2000, considering first and second ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first or second ancestry. By construction, the sum of these shares exceed 100 percent. In the case of the United States, both American and Native American ancestries are attributed to it.

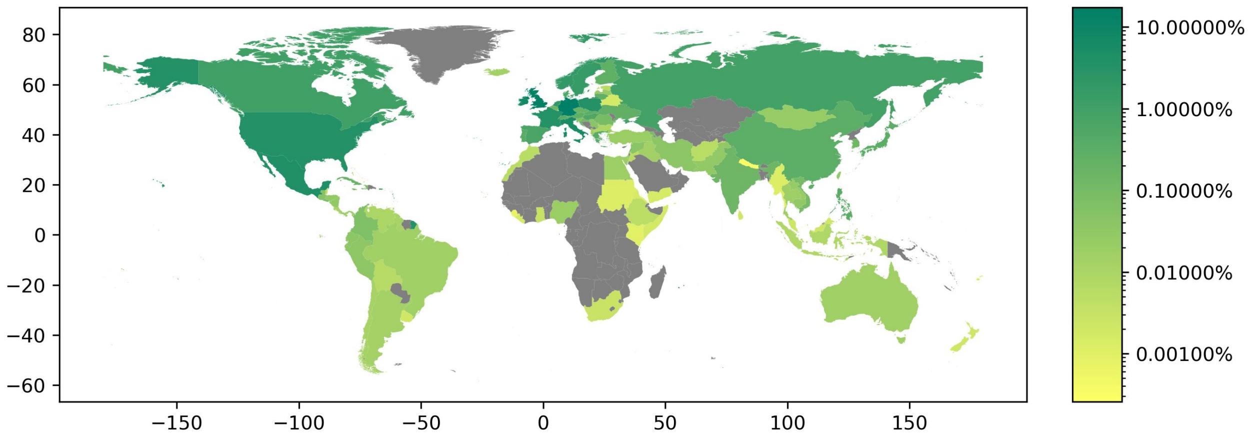
Figure 3: Proportional Distribution of Native Ancestry by Country in 2000 (logarithmic scale)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first ancestry. In the case of the United States, both American and Native American ancestries are attributed to it.

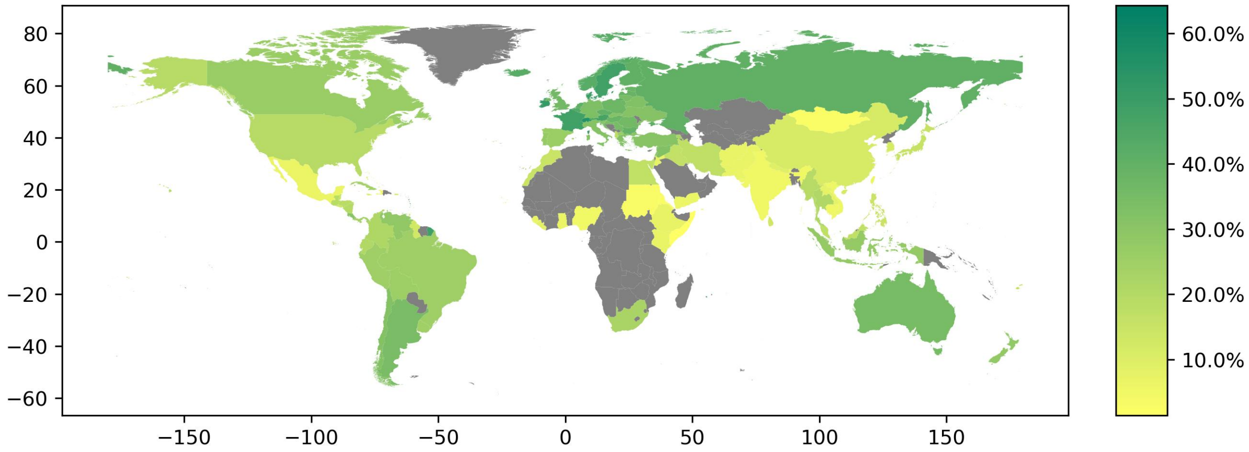
Figure 4: Proportional Distribution of Native Ancestry by Country in 2000, considering first and second ancestry (logarithmic scale)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first or second ancestry. By construction, the sum of these shares exceed 100 percent. In the case of the United States, both American and Native American ancestries are attributed to it.

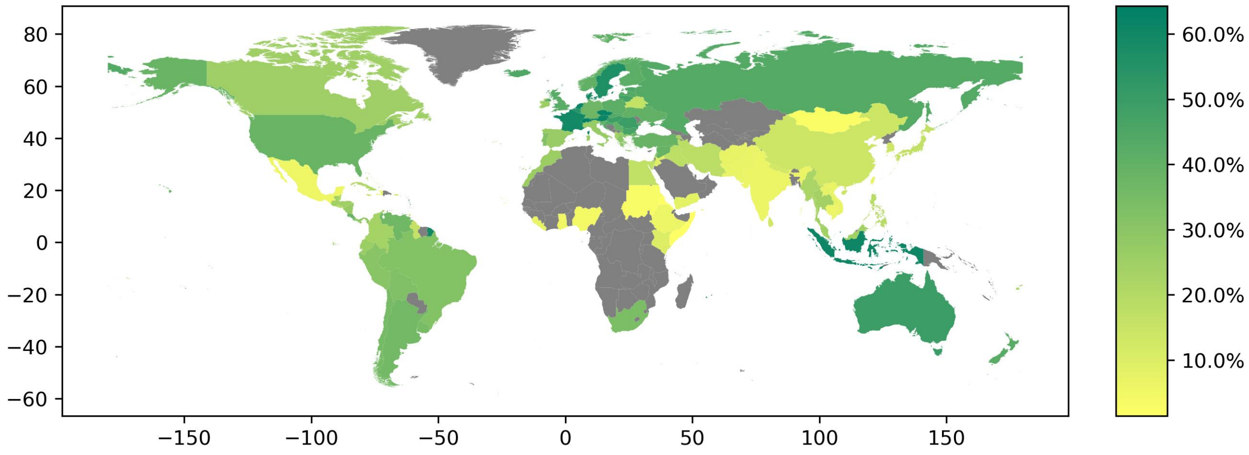
Figure 5: Incidence of multiple ancestries



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: We examine the percentage of individuals with a specific ancestral background who identify as having a secondary ancestry. For example, we analyze the proportion of individuals with Italian as their primary ancestry who also report having a secondary ancestry.

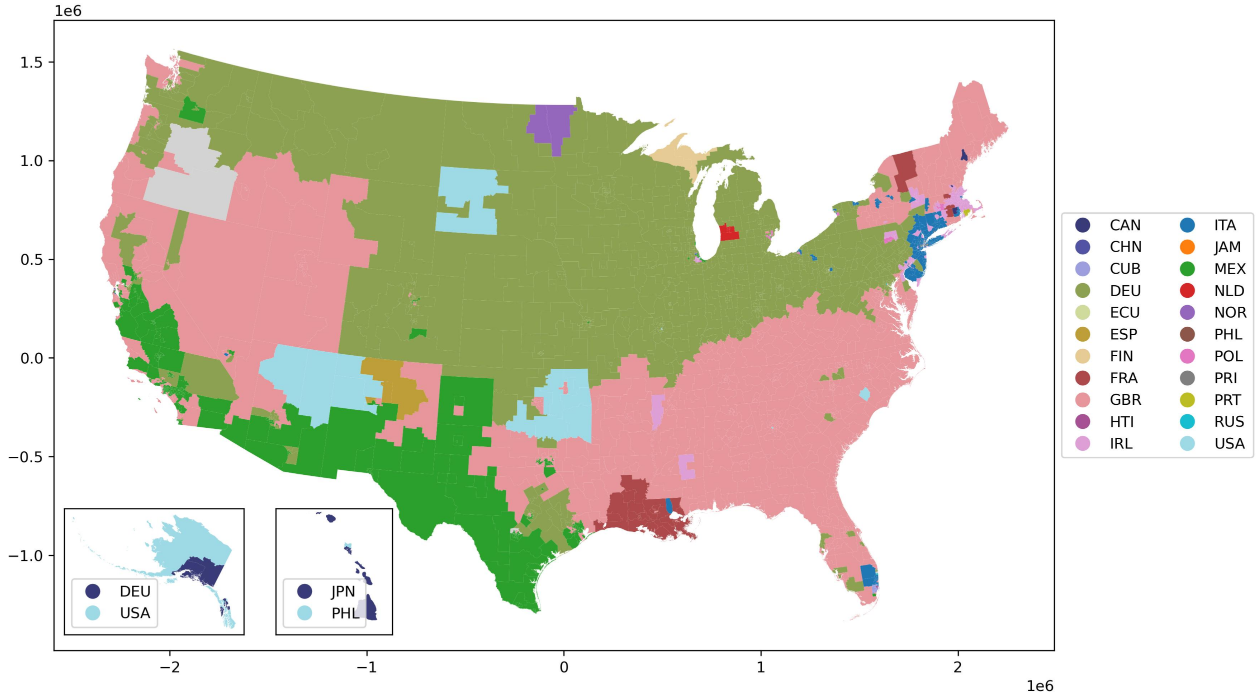
Figure 6: Incidence of multiple ancestries
(first and second ancestry)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: for each ancestry, we compute the share of natives who also report a second ancestry, e.g., we compute the share of natives with Italian first or second ancestry that also report another ancestry.

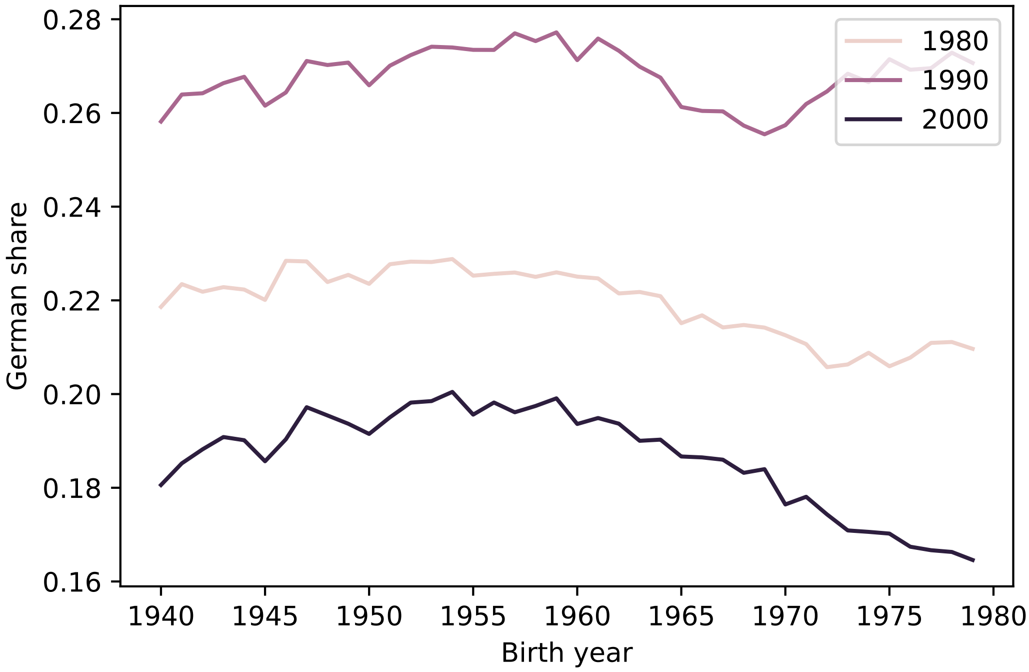
Figure 7: Most prevalent ancestry in each PUMA



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

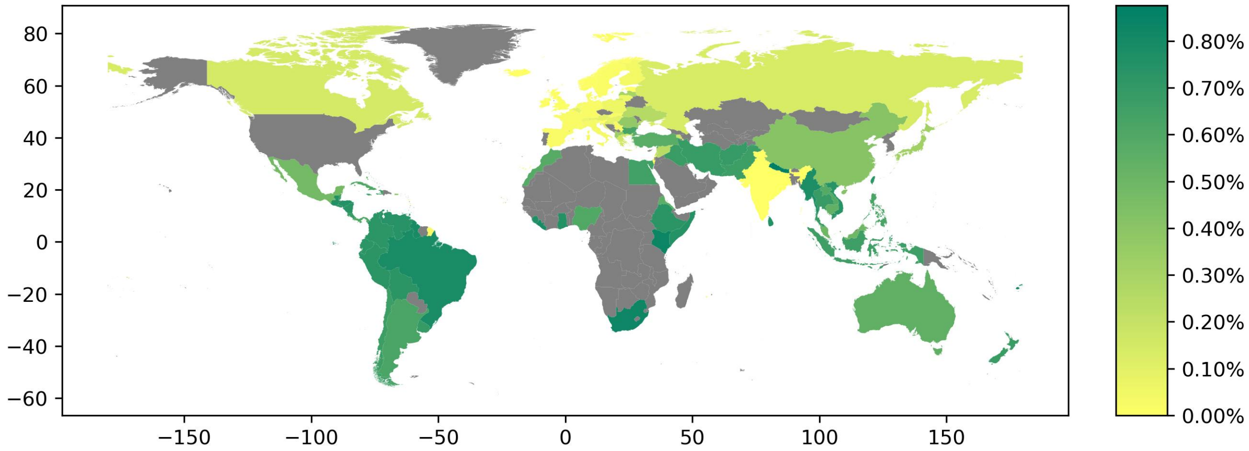
Notes: For each Public Use Microdata Area (PUMA), we identify and assign the most prevalent ancestral heritage among native residents.

Figure 8: Natives with Germany ancestry by birth cohort, different census years



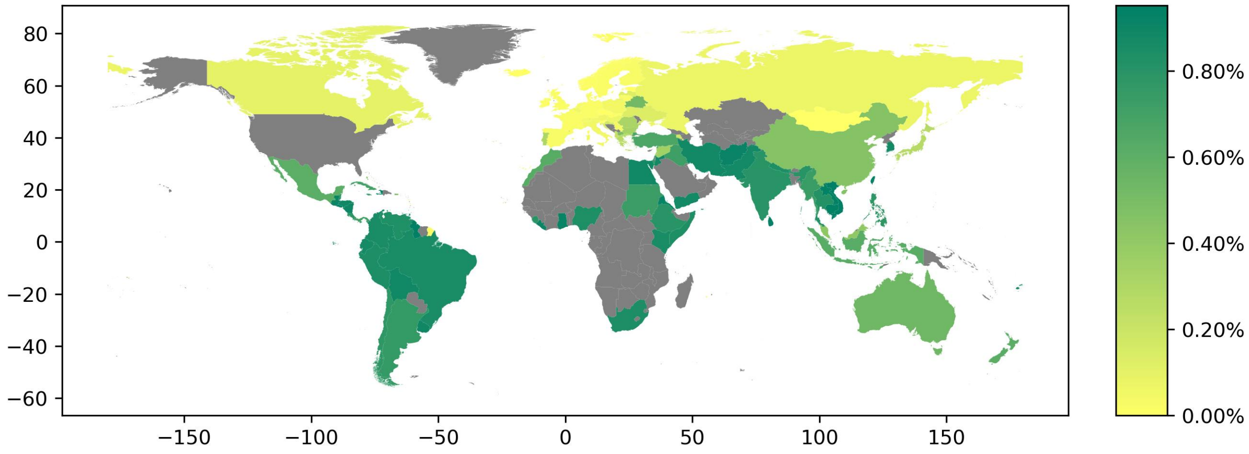
Data sources: Authors' elaboration on the 1980, 1990 and 2000 Census (Ruggles et al., 2023).

Figure 9: Share of individuals born in the ancestral country



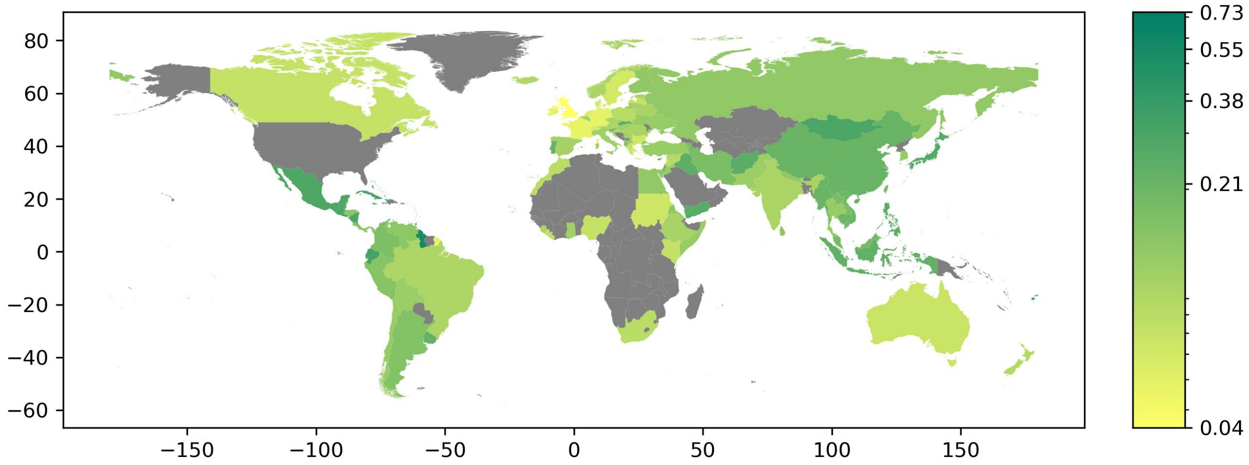
Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Figure 10: Share of natives co-residing with at least one parent whose first ancestry coincides with the maternal or paternal country of birth



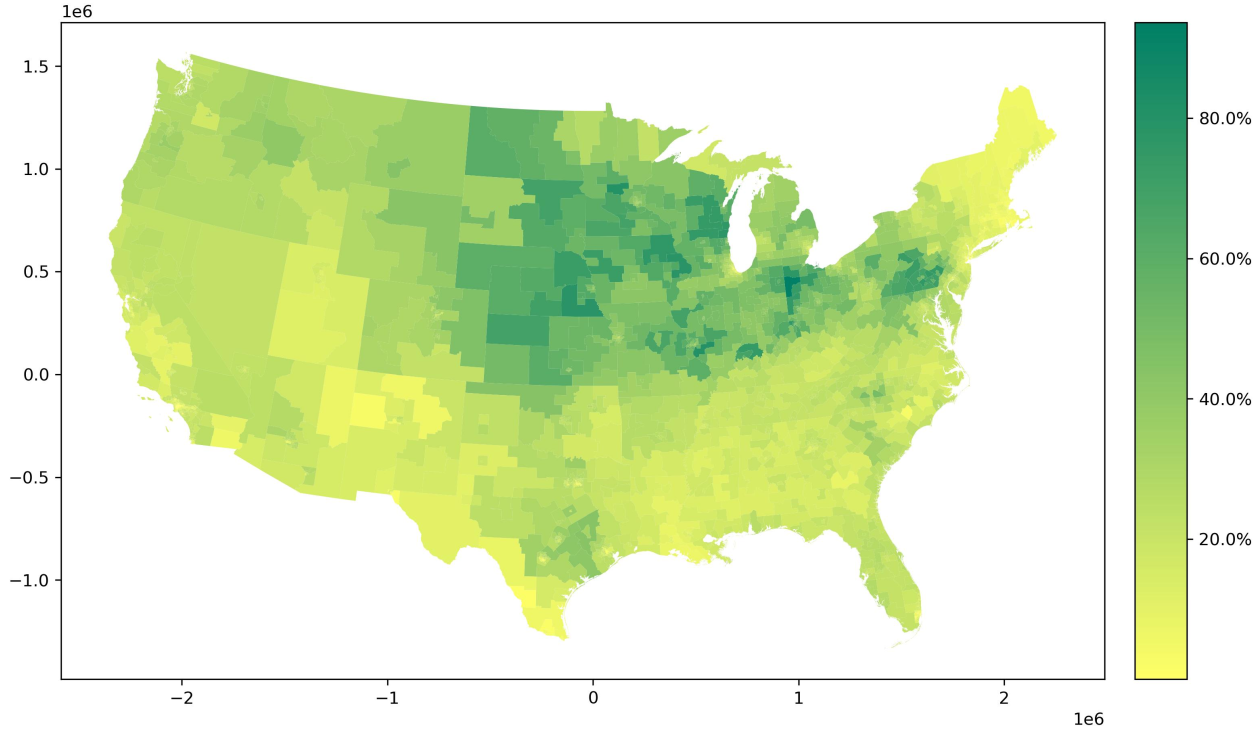
Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Figure 11: Herfindahl-Hirschman Index of spatial concentration of natives of foreign ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

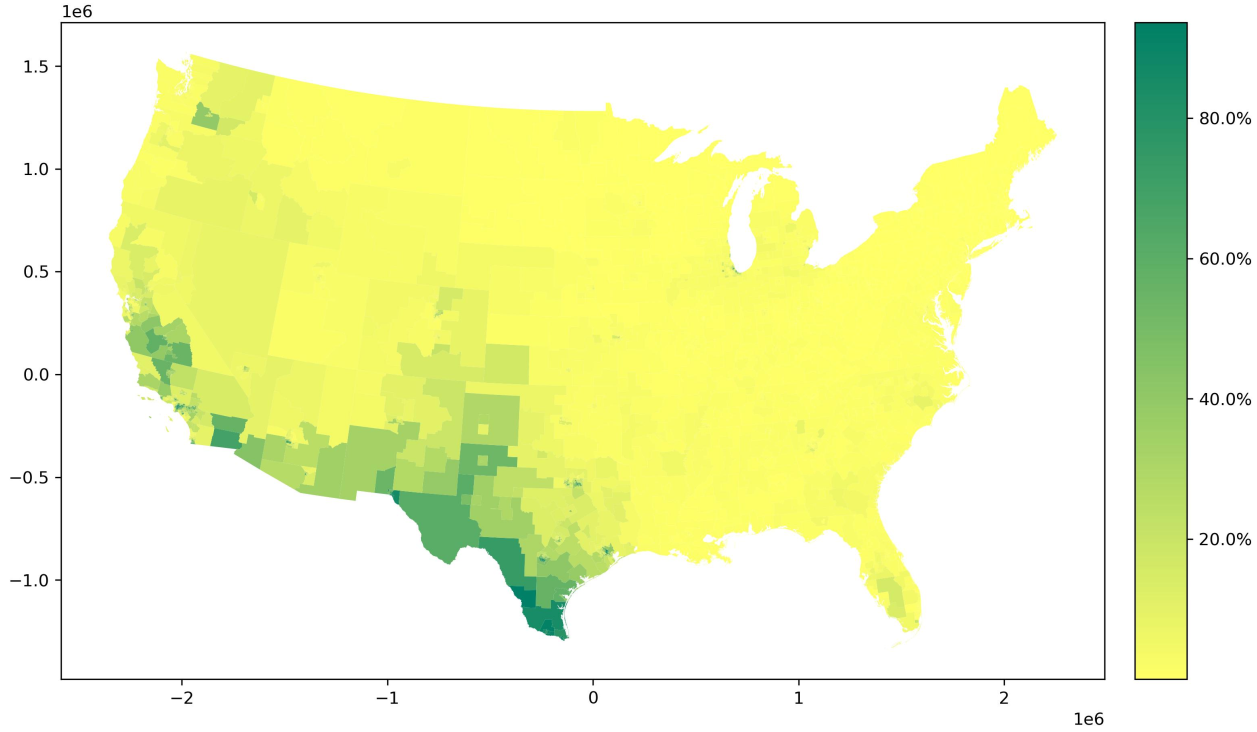
Figure 12: German ancestry among natives in each PUMA



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure present the percentage of native-born population that identify with the German ancestry as first ancestry. The sample is composed of individuals born in the USA that report at least one ancestry.

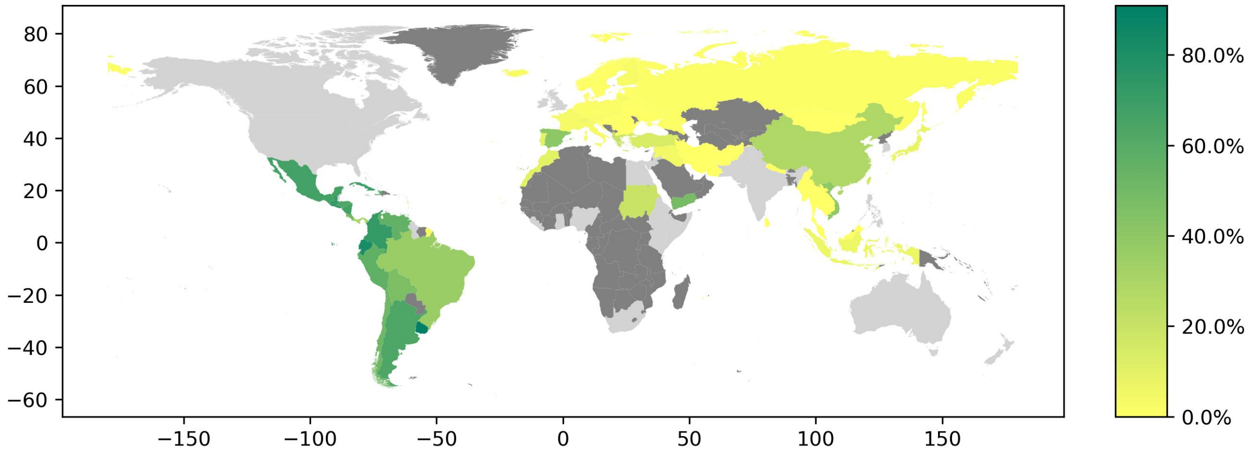
Figure 13: Mexican ancestry among natives in each PUMA



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure present the percentage of native-born population that identify with the Mexican ancestry as first ancestry. The sample is composed of individuals born in the USA that report at least one ancestry.

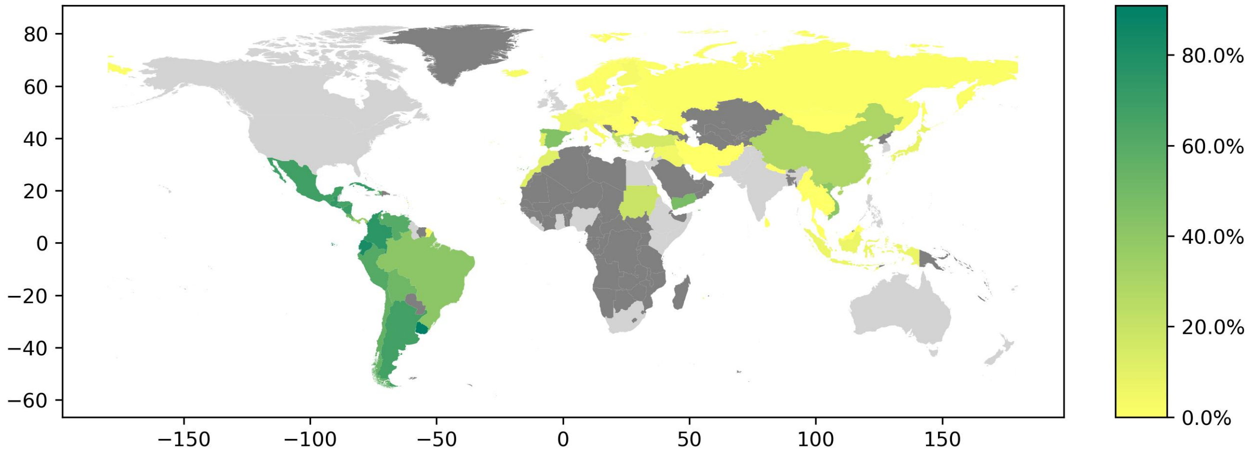
Figure 14: Share of natives speaking the ancestral language, by country of ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure displays the percentage of individuals who claim ancestry in a particular country (first ancestry only) and also speak an official language of that country. The analysis includes only individuals with at least one reported ancestry, and countries where English is an official language are represented in light gray.

Figure 15: Share of natives speaking the ancestral language, by country of ancestry (first and second ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure displays the percentage of individuals who claim ancestry in a particular country (first or second ancestry) and also speak an official language of that country. The analysis includes only individuals with at least one reported ancestry, and countries where English is an official language are represented in light gray.