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

Simone Bertoli, Melchior Clerc, Jordan Loper and Èric Roca Fernández

2024 Deep-Rooted Factors in Comparative Development

Introduction

- Why is income inequality higher in some groups?
- What explains risk aversion and long-term orientation?
- Modern outcomes and cultural norms reflect the characteristics of a distant past.
    - Geographic,
    - climatic,
    - and genetic factors.

Introduction

- To establish a causal effect, the literature has turned to using the *epidemiological approach* (Fernández, 2007).
    - Compare individuals with *different* origins
    - who reside in a *single* country.
        - Rationale for *multiple* origins: Exploiting differences across origins in cultural, geographic, climatic and genetic factors.
        - Rationale for *single* country: exposure to the same institutional framework, labour market, incentives, etc.

The epidemiological approach: biases

- The literature has emphasized an *attenuation bias* in the epidemiological approach (Fernández, 2011).
  - Multiple cultural sources: parents are not the only source of cultural transmission, potentially making the cultural proxy insignificant.
  - Selective migration: migrants moving to a given destination from different countries of origin are likely to be more similar than the populations at origin are.

The epidemiological approach: biases

- We propose a novel potential bias in the epidemiological approach in the form of endogeneity.
- Depending on the specification, it may overestimate the true causal effect of an origin-specific variable.

- The equation typically used in the epidemiological approach is:

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

- y_{iok} outcome for individual i of origin o, residing in location k
- w_o is the origin-specific variable of interest
- $\boldsymbol{x}_o$ and $\boldsymbol{x}_i$ origin-level and individual-level variables; d_k represents area of residence FE
- t_i is the time elapsed since ancestral migration for individual i

The epidemiological approach: biases

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

- The estimated value of α can be biased if:
  - corr(w_o, t_i) $\neq 0$ and $\lambda \neq 0$
- Example:
  - Crop yield at origin (w_o) determines future-orientation. To test this, a researcher regresses having a private pension plan (y_{iok}) on w_o. However, origins that were more future-oriented may have started migrating earlier (t_i). By having spend longer in the receiving country, they have better knowledge about the institutional aspects of public and private pension systems.

This paper

- Focuses on the US Census (central focal point in the epidemiological approach).
- Indicates possible variables to relate an individual with a foreign ancestry.
- Describes heterogeneities across individuals with different origins.
    - The dimensions of heterogeneity can be used to *control* for time since migration.

Foreign origins in the US Census

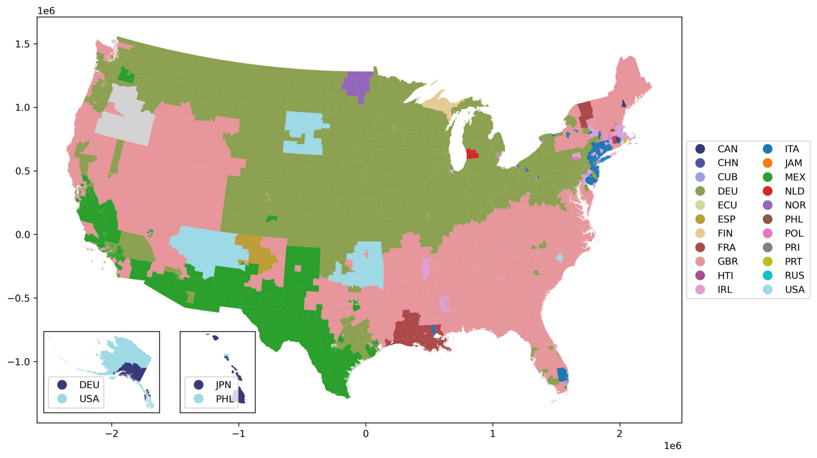
- Variables relating an individual with a foreign ancestry
  - Birth country
    - Only available for first-generation migrants, self-selection into migration.
  - Parental country of origin
    - Available until the 1970 census and in the CPS March supplement.
    - Can be constructed for second-generation migrants *co-habiting* with their parents.
  - Language spoken
    - Difficult to assign a country, self-selection.
  - Ancestry
    - “What is this person’s ancestry or ethnic origin?”; Census (since 1980), ACS, General Social Survey.

Census question from 1980 to 2000

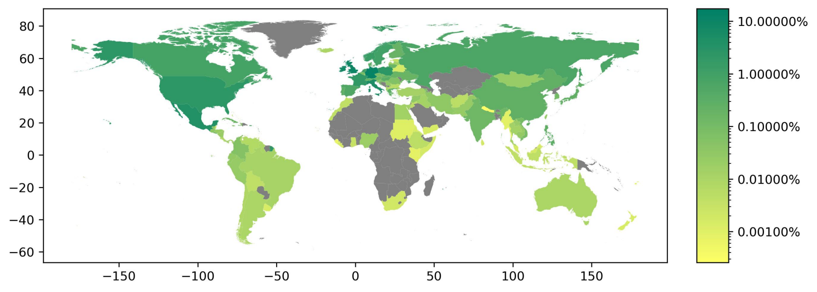
From country of birth to ancestry

- The literature has progressively shifted from a focus on *first-generation immigrants* (country of birth), to the analysis of *natives of immigrant origin* (country of birth and ancestry).
    - Papers relying on ancestry: Antecol (2000), Guiso et al. (2006), Fernández and Fogli (2006), Alesina and Giuliano (2011), Alesina et al. (2015), Galor and Özak (2016), Giavazzi et al., (2019), Arbatlı et al. (2020), Giuliano and Nunn (2021), Galor et al. (2023).
- Rationale: influence of non-random selection into migration should be diluted over time
- BUT influence of unobserved heterogeneities (e.g. time since migration) should be stronger

Main ancestry in each PUMA among US natives in 2000



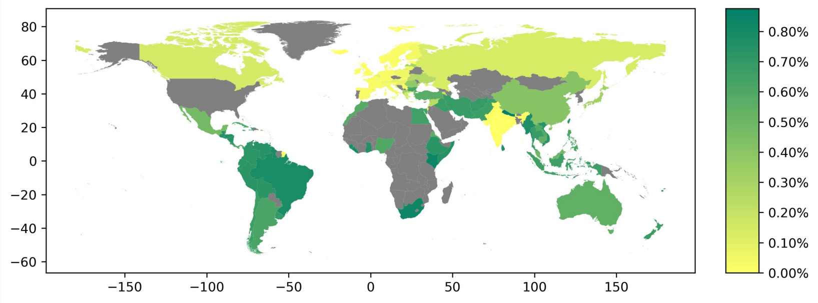
Distribution of ancestries among natives in 2000



Building proxies of overlooked confounders

- Heterogeneities across individuals of different origins are overlooked in the literature and may confound epidemiological estimations:
  1. time since ancestral migration
     - % of individuals born in the ancestral country
  2. spatial concentration
     - Migrants tend to choose locations based on pre-existing networks
  3. attachment to origin's identity
     - speaking the ancestral language (Giuliano and Nunn, 2021)
- We build proxies for each of these dimensions (using data from the 2000 census, identifying 109 distinct ancestral countries)

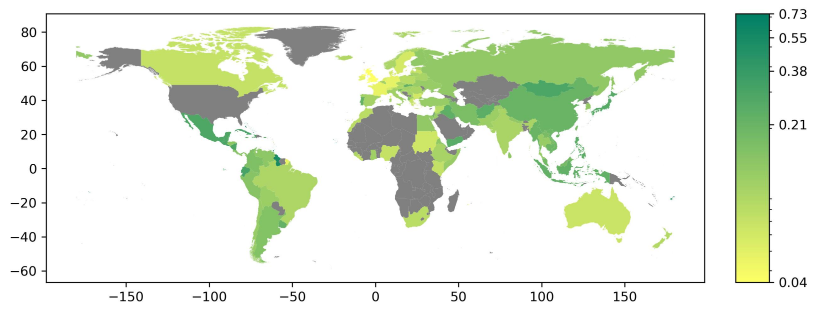
Share of first-generation immigrants (v_o^1)



v_o^1 = Share of the population (natives and immigrants) reporting an ancestry in country o that is born in the ancestral country

- A higher share of first-generation immigrants (v_o^1): a shorter average time since ancestral migration
- $v_{DEU}^1 = 0.021$ vs. $v_{MEX}^1 = 0.472$; mean $= 0.448$

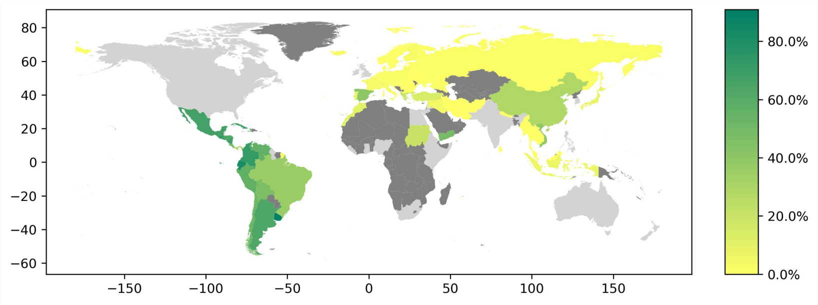
Spatial concentration (v_o^2)



v_o^2 = HHI of spatial concentration of the population (natives only) reporting an ancestry in country *o* across states

- A higher spatial concentration (v_o^2): a shorter average time since ancestral migration (but local labor demand shocks also matter)
- $v_{DEU}^2 = 0.048$ vs. $v_{MEX}^2 = 0.283$; mean = 0.182

Share of natives speaking their ancestral language (v_o^3)



v_o^3 = Share of the population reporting an ancestry in (non-English speaking) country o that speak their ancestral language at home

- A higher share (v_o^3): a stronger attachment to origin's identity (negatively correlated with time since migration too) (Giuliano and Nunn, 2021)
- $v_{DEU}^3 = 0.015$ vs. $v_{MEX}^3 = 0.557$; mean $= 0.247$

Validity of our proxies

- Our proxies built on the basis of self-reported ancestry are coherent with what we know about the history of origin-specific migration flows to the United States:
  - 8 out of 10 origin countries with lowest values of v_o^1, v_o^2 or v_o^3 are European (migration to the US in a more distant past)
  - Latin American countries tend to be among the ones with the highest values of these variables (more recent waves of migration to the US)
- Our proxies are well correlated to each other:
  - $\text{corr}(v_o^1; v_o^2) = 0.863$,
  - $\text{corr}(v_o^1; v_o^3) = 0.402$,
  - $\text{corr}(v_o^2; v_o^3) = 0.500$

Variability of our proxies

- R^2 of simple (weighted) regression of our proxies continent dummies:
    - v_o^1: 0.804
    - v_o^2: 0.700
    - v_o^3: 0.101
- Most of the variability in v_o^1 and v_o^2 is across rather than within continent.
- Takeaway: including continental dummies absorbs most of the variation in time since ancestral migration. This is not the case for attachment to the foreign ancestry.

Selection into unobservables

- Following Oster (2019), explore how the R^2 of several regressions changes when including our proxies.
  - Baseline regression with individual-level controls, provides R^2_{min}
  - Augment it with a full set of origin fixed effects, provides the maximum R^2_{max} that any origin-specific variable can have
  - Replace the origin fixed effects with our proxies and compare how far the R^2 is from R^2_{max}
- R^2 is relatively close to R^2_{max}, this indicates that, for the outcome y, most of the variation attributable to any origin-specific variable w_o comes mostly from differences in time since migration.

Selection into unobservables

	Inter-ethnic marriage		Education level	
	(1)	(2)	(3)	(4)
Age	−0.001	−0.001	0.494	0.514
	(0.000)***	(0.000)***	(0.004)***	(0.004)***
Age sq.	0.000	0.000	−0.007	−0.007
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Gender	−0.011		−0.558	
	(0.000)***		(0.021)***	
Speaks trad. lang.	−0.037		−1.420	
	(0.000)***		(0.023)***	
% born in orig.	0.477		−7.308	
	(0.002)***		(0.188)***	
Herfindahl index	0.735		−26.147	
	(0.006)***		(0.387)***	
R^2_{min}	0.045	0.045	0.154	0.154
R^2_{max}	0.100	0.100	0.180	0.180
R^2	0.082		0.165	
R^2		0.080		0.165
Observations	6582979	6579946	4262914	4262612
Continent FE	No	Yes	No	Yes
N. countries	105	105	105	105

Concluding remarks

- History matters: natives of different foreign ancestries greatly differ in terms of time since ancestral migration.
- Potential for endogeneity.
- Controlling for the proxies we propose: bias-reducing strategy

Appendix

Census question from 1980 to 2000

“What is this person’s ancestry or ethnic origin?”

- Respondents can report a country, or what the Census Bureau defines a “general heritage” (e.g., African, European).
- Multiple answers are possible (up to three in 1980, two since 1990).
- American ancestry recorded only if no other ancestry is mentioned.

Back Next

Census question from 1980 to 2000

- The share of the population (natives and immigrants) not reporting an ancestry stood at around 10 percent in 1980 and 1990, and it then jumped to 19 percent in 1990.
- Around 40% of the individuals reporting one ancestry report multiple ancestries.
- Major variations in the share of respondents reporting a given foreign ancestry.

Back Next

Census question from 1980 to 2000 (cont'd)

- The share of the population with an English ancestry collapsed from 26.3 to 16.1 percent between 1980 and 1990 (Rosenwaike, 1993), when Germany became the first ancestry. Why?
  - In 1980, the ancestry question followed a question on English proficiency.
  - 1980 Census: "What is this person's ancestry?" (For example: Afro-Amer., *English*, French, *German*, [...]).
  - In 1990, the ancestry question came *before* the question on English proficiency, and English was *not* included among the examples.
  - 1990 Census: 'What is this person's ancestry or ethnic origin?" (For example: *German*, Italian, Afro-Amer., [...]).
  - German was *not* listed among the examples in the 2000 Census.

Back