README

to "How Cultural Diversity Drives Innovation: Surnames and Patents in U.S. History"

August 25, 2025

The folder contains data files and all code needed to replicate the tables and figures of the paper and online appendix. The scripts are located in the ./code/ folder. Several RData files containing the data needed to replicate the results are located in the ./data/ folder.

We are unable to publicly post some of the data used in the analysis. First, we gained access to individual-level 1850–1940 Population Census data including surnames via an agreement with the Minnesota Population Center. Other researchers can apply for access to the restricted version of the data, as outlined here [link accessed August 15, 2025]. Second, we obtained access to the Comprehensive Universe of U.S. Patents (CUSP) compiled by Enrico Berkes. Researchers can contact Enrico for access to this data set [link accessed August 15, 2025].

Required software

- R – We used version 4.5.0 (2025-04-11)
- Stata – We used 19.5 MP

Paths configuration Edit the paths in config.json:

- Project folder – Absolute path to where your replication folder is located
- Stata bin – Absolute path to where your Stata binary is located
- Stata version – The version of your Stata

Code

- ./code/1stataSetup.do – Install Stata user-written programs.
- ./code/2RSetup.r – Install R packages, define paths and functions.
- ./code/3analysis.r – Calls scripts to generate all tables and figures of the paper and the appendix.

Output All replicated tables and figures are saved in the ./output/ folder. Tables are saved as *.tex files, and figures as *.eps or *.pdf files.

Data The following data files are needed to reproduce the figures and tables in the paper and online appendix. I flag restricted-access data with red color text and an * bullet point.

* ivSurnameStocksMetaphone1900BoundariesPanel18951940.csv – the surname-county-level data set used in Table 1 and Appendix Figures B1.
- countyLevel19001940.csv – the main county-level data set used in Tables 2, 5, 8, 9 and Figure 1, and in Appendix Tables B1, B3-B7, B10, B12, B13, B16, B18, B21, B24, B27, B30, B34-B36, B39, B42, B43, B45, B52, B53 and Appendix Figures B1, B6, and B9-B12.
* surnameCountyLevel.csv – the surname-county-level data set used in Table 3, and in Appendix Tables B2, B8, B11, B14, B17, B19, B22, B25, B28, B31, B33, B34, B37, B40 and Appendix Figure B14.
* inventorLevel.csv – the inventor-level data set used in Table 4, and in Appendix Tables B9 and B12, B15, B20, B23, B26, B29, B32, B35, B38, B41, and B54.
* patentLevel.csv – the patent-level data set used in Table 6 and Appendix Tables B46-B50.
* 1940.csv and 1880.csv – the 1940 and 1880 censuses used in Table 7 and Appendix Table B51.
* cusp_patents.csv and patents_fyear_iyear.csv – the CUSP patent data sets used in Table 7, and in Appendix Table B51 and Appendix Figure B15.
* cusp_patents_inventor_surnames_fips.csv – the surnames of inventors in the CUSP patent data set used in Table 7, and in Appendix Table B51 and Appendix Figure B15.
* geneticDistanceNameLevelMetaphone.RData – the surname-ancestral-country cross-walk used in Table 7 and Appendix Table B51.
- US_tl2000_cnty_shp_1790_2000.Rdata and US_tl2000_state_shp_1790_2000.Rdata – the shapefiles used in Appendix Figures B2, B3, and B6-B10.
- countyLevel18501940.csv – the county-level data set used in Appendix Figures B2-B4, B7-B9, and B13.

* immigrationNameLevel1900BoundariesPanel18801930.csv – the surname-level immigration data used in Appendix Figures B5 and B6.
* ivAncestralCountryStocksMetaphone1900BoundariesPanel18951940.csv – the ancestral-country-level immigration data used in Appendix Table B44.