**README File for Replication Package**

**From Immigrants to Americans: Race and Assimilation during the Great Migration**

**Vasiliki Fouka & Soumyajit Mazumder & Marco Tabellini**

This replication package includes

* Data: datasets (.dta) to replicate all analyses in the main text
* Codes: STATA files (.do) to replicate results reported in all figures and tables of the main text
* Results: outputs for all figures (.eps) and tables (.xls) displayed in the main text

**Dataset list**

In order to replicate all figures and tables of the main text, the following 5 datasets are included:

1. **main\_dataset:** it includes data needed to produce most of the results in the main paper (Figures 1, 3, 5, 6, 7 and Tables 1, 2, 3, 6, 7).
2. **linked\_sample**: it includes data needed to produce results for the linked, panel dataset of individuals (col. 6 of Table 2, Panel B of Table 3)
3. **secgen\_names**: it includes data needed to produce results on foreign name index (col. 4 of Table 3)
4. **segregation\_msa**: it includes data needed to produce results on residential segregation at the MSA-national origin level (cols. 5 and 6 of Table 3)
5. **newspapers\_msa**: it includes data needed to produce results on historical press at the MSA level (Figure 4 and Tables 4, 5)

**Data availability and provenance statements**

The paper uses data extracted from three data sources:

1) Integrated Public Use Microdata Series – IPUMS USA (public use data)

This dataset consists of a series of compatible-format individual-level representative samples of the United States for the years 1900, 1910, 1920, and 1930. In particular, we make use of complete count data. All data are coded and anonymous. The variable [VERSIONHIST](https://usa.ipums.org/usa-action/variables/VERSIONHIST) identifies the public-release version of a dataset.

Data are free and publicly available. They can be downloaded via the IPUMS USA extract system upon registration (see the link: <https://usa.ipums.org/usa-action/samples>). The extraction system creates a record layout tailored to the needs of each user and the approval is instantaneous.

These data are cited as: *Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek, Integrated Public Use Microdata Series: Version 6.0 [Machine-Readable Database], University of Minnesota, 2015.*

2) Integrated Public Use Microdata Series – IPUMS USA (restricted access data)

We complement data taken from [1] using variables present in the IPUMS restricted access dataset. From this database, we obtain respondents’ names and their position on original Census manuscripts. To get access to the restricted use data, researchers need to find a sponsor within the National Bureau of Economic Research (NBER), and fill out an application request, which must be sent to the following email address: [ipums@umn.edu](mailto:ipums@umn.edu). Upon approval, researchers can access the NBER server, where the restricted use data are stored.

For confidentiality reasons, individual level data containing restricted use variables (including those used in our paper) cannot be downloaded from the server. Thus, we created the variables needed for the analysis on the NBER server. We then exported the de-anonymized data, which were merged into the public use IPUMS data [1] using the variable HISTID – a variable present in both the restricted and the public use datasets.

These data are cited as: *Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek, Integrated Public Use Microdata Series: Version 6.0 [Machine-Readable Database], University of Minnesota, 2015.*

3) Newspapers.com

We retrieve words used in historical newspapers from the website <https://www.newspapers.com/>, last accessed in August 2020. This is the largest online newspaper archive consisting of more than 641 million pages of historical newspapers from 20,700 newspapers from around the United States and beyond. The online archive is accessible upon registration.

There is no specific citation for this dataset, except the website itself (i.e., <https://www.newspapers.com/>).

**Data sources**

The datasets in the replication folder are obtained from the sources listed above as follows:

a) **main\_dataset:** The raw data are from IPUMS USA, public use data [1]. The dataset available in the replication package is obtained from this source by restricting the sample to the 108 MSAs outside the US South (the complete list of MSAs in our sample is reported in Table A.1) for the three Census decades 1910, 1920, 1930.

b) **linked\_sample:** The raw data are from IPUMS USA, restricted use data [2]. The dataset available in the replication package is obtained from this source as follows. We start from the universe of foreign-born men in 1910 who are unique by first and last name, birthplace and year of birth and we restrict the sample to foreign-born men in our main dataset as defined above in point (a). The variable that uniquely identifies individuals is HISTID. Following standard automated census-linking procedures used in the economic history literature (Ferrie, 1996; Abramitzky et al., 2014), we construct a panel of immigrants linked across census years. The construction of the sample is discussed in Section C.2 of the paper’s appendix.

As noted above, name fields are not publicly available. However, our final dataset can be redistributed, since it does not contain any variable of the restricted use data.

c) **secgen\_names**: The raw data are from IPUMS USA, restricted use data [2]. The dataset available in the replication package is obtained from this source by restricting our sample to US-born children of a foreign-born father for the three Census decades 1910, 1920, 1930. To construct the American sounding name index, discussed in more details in Section C.1 of the paper’s appendix, we use the name (namefrst and namelast) fields available through a special arrangement with IPUMS.

As noted above, name fields are not publicly available. However, our final dataset can be redistributed, since it does not contain any variable of the restricted use data.

d) **segregation\_msa:** The raw data are from IPUMS USA, restricted use data [2]. The dataset available in our replication package is obtained from this source by restricting our sample to households heads for the three Census decades 1910, 1920, 1930. In order to identify neighbors and construct the variable described in Section C.1 of the paper’s appendix, we rely on three variables (reel, pageno, line).

As noted above, these data are not publicly available. However, our final dataset can be redistributed, since it does not contain any variable of the restricted use data.

e) **newspapers\_msa:** The raw data are from Newspapers.com [3]. The dataset available in the replication package is obtained from this source as follows. First, we scraped all pages containing the words included in the analysis conducted in the paper. This step was implemented focusing on the time span (1900 to 1930) and the geography (the 108 non-southern MSAs) of the paper. The website Newspapers.com [3] was last accessed in August 2020. Then, the yearly-MSA dataset was collapsed to the decade-MSA level (summing the frequencies of individual words across all years within each decade). The dataset available in the replication package is thus at the decade-MSA level.

### Details on each Data Source

All the datasets are in .dta format, i.e. they are all in STATA format. Data from [1] and [2] were downloaded in 2018. Newspapers data [3] were last downloaded in 2020.

### Summary of Availability

As explained in detail above, some of the data **cannot be made** publicly available.

**Statement about Rights**

The author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

**Description of programs/code and instructions for replication**

Results reported in the paper were produced using STATA 16. In order to replicate results, please make reference to the master dofile. First, set the working directory to the replication folder; then, codes should execute. There is one .do file for each table/figure (named accordingly).

In order to run the replication, you may need to install the following packages from the SSC archive: reghdfe, xtivreg2, binscatter, ceofplot, asdoc, ivreghdfe, ivreg2, xtivreg2, ranktest

**List of codes**

|  |  |  |
| --- | --- | --- |
| **Figure/Table #** | **Program** | **Output File** |
| Figure 1 | Figure1.do | Figure1a.eps, Figure1b.eps |
| Figure 3 | Figure3.do | Figure3.eps |
| Figure 4 | Figure4.do | Figure4a.eps, Figure4b.eps |
| Figure 5 | Figure5.do | Figure5a.eps, Figure5b.eps, Figure5c.eps, Figure5d.eps |
| Figure 6 | Figure6.do | Figure6a.eps, Figure6b.eps |
| Figure 7 | Figure7.do | Figure7a.eps, Figure7b.eps |
| Table 1 | Table1.do | Table1.doc |
| Table 2 | Table2.do | Table2\_A,.xls Table2\_B.xls |
| Table 3 | Table3.do | Table3\_A,.xls Table3\_B.xls |
| Table 4 | Table4.do | Table4.xls |
| Table 5 | Table5.do | Table5\_A.xls, Table5\_B.xls |
| Table 6 | Table6.do | Table6.xls |
| Table 7 | Table7.do | Table7.xls |

**References**

Abramitzky, Ran and Leah Platt Boustan, and Myera Rashid, Census Linking Project: Version 1.0 [dataset], https://censuslinkingproject.org, 2020

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson, “A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration,” Journal of Political Economy,2014,122(3), 467–506

Ferrie, Joseph, “A New Sample of Males Linked from the Public Use Micro Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules,” Historical Methods, 1996,29(4), 141–156.

Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek, Integrated Public Use Microdata Series: Version 6.0 [Machine-Readable Database], University of Minnesota, 2015.