Read Me File for the Census Tree Dataset

Contact information:
Joseph Price joseph_price@byu.edu
Kasey Buckles kbuckles@nd.edu
Adrian Haws aah226@cornell.edu
Haley Wilbert hwilbert@nd.edu

The Crosswalks

Each .csv file downloaded from the OpenICSPR repository contains a crosswalk that can be used to link records between the indicated years, using the IPUMS versions of the full count U.S. censuses from 1850 to 1940. For example, the file "1900_1910.csv" has the variables *histid1900* and *histid1910* that can be merged to the 1900 and 1910 IPUMS census files to create an individual-level panel dataset of men and women who appear in both censuses.

Each file contains the following variables:

histidYYYY	contains the unique IPUMS identifier for each year (two variables)
clp	=1 if link is included in Census Linking Project
mlp	=1 if link is included in IPUMS Multigenerational Longitudinal Panel
xgb	=1 if link is generated by XGBoost ML algorithm
family_tree	=1 if link is on Family Tree (i.e. identified by FamilySearch users)
direct_hint	=1 if link is included in FamilySearch census-to-census hints
profile_hint	=1 if link is included in FamilySearch census-to-profile hints
implied	=1 if link is implied by links established between other census pairs

For more information on the link sources and on the creation of the Census Tree, see Buckles, Haws, Price, and Wilbert (2023), available at https://censustree.org.

Step-by-Step Instructions for Using the Census Tree Data

1. Download the data from the OpenICPSR repository, and unzip the file if necessary.
2. Download the IPUMS versions of the relevant censuses, available here. (The crosswalks will also work with the restricted-access versions of the IPUMS censuses.)
3. Prepare the IPUMS files for merging; we have provided sample code below to suggest one way of doing this in Stata.
4. Use the histid variable from each census year to merge them both to the crosswalk.
5. Be sure to properly cite any data used in your work, and tell us about it! Citation instructions can be found at https://censustree.org/overview

Using the Data in Stata

The following code may be helpful for using the crosswalks in Stata, using the 1900-1910 crosswalk as an example.

Opening the .csv file in Stata:

```
import delimited using "filelocation\1900_1910.csv", varnames(1)
```

Labeling the variables:

```
label var clp "=1 if link is included in Census Linking Project"
label var mlp "=1 if link is included in IPUMS Multigenerational Longitudinal
Panel"
label var xgb "=1 if link is generated by XGBoost ML algorithm"
label var family_tree "=1 if link is on Family Tree (i.e. identified by
FamilySearch users)"
label var direct_hint "=1 if link is included in FamilySearch census-to-
census hints"
label var profile_hint  "=1 if link is included in FamilySearch census-to-
profile hints"
label var implied "=1 if link is implied by links established between other
census pairs"
```

Preparing the IPUMS censuses for merge (assuming all necessary files are in the same file location, and original IPUMS files are named YYYY_full.dta):

```
use "filelocation\1900_full.dta", clear
* keep only the variables you need to make smaller files for easier merge—for
* example, suppose you only need to know age and sex
keep vars histid age sex
* rename the variables to indicate the year they are from
for var histid age sex: rename X X1900
save "filelocation\1900_merge.dta", replace

use "filelocation\1910_full.dta", clear
* keep only the variables you need to make smaller files for easier merge—for
* example, suppose you only need to know age and sex
keep vars histid age sex
* rename the variables to indicate the year they are from
for var histid age sex: rename X X1910
save "filelocation\1910_merge.dta", replace
```

Merging the IPUMS censuses to the crosswalks (assuming all necessary files are in the same file location, and the prepared IPUMS files are named YYYY_merge.dta):

```
import delimited using "filelocation\1900_1910.csv", varnames(1)
** the next two starred lines of code may be necessary if your IPUMS files
** have histids with lowercase letters
* replace histid1900 = strlower(histid1900)
```

```
* replace histid1910 = strlower(histid1910)
merge 1:1 histid1900 using "filelocation\1900_merge.dta", keep(3)
merge 1:1 histid1910 using "filelocation\1910_merge.dta", keep(3)
```

Other Resources Available at https://censustree.org

- Quick Start Guide
- FAQ page with information about the Census Tree
- Link to Buckles, Haws, Price, and Wilbert (2023) with detailed description of how the Census Tree is created
- Links to access the OpenICPSR repositories for all year-to-year crosswalks
- Information on how to cite
- “Contact Us” form for telling us about your work using the Census Tree.