How Matched Crosswalks were Created

Code that runs each of the following sections is contained in *create_matched_crosswalks.do*. This document outlines how the matched crosswalks were created and explains each section in more detail. The code is based on the method developed by Abramitzky, Boustan and Eriksson (2012, 2014, 2017, 2020). For details on the ABE matching methods used in this code, please visit https://ranabr.people.stanford.edu/matching-codes

Section 0: Set Up

The set-up part of the code allows users to pick which two Census years they would like to match as well as the methods they would like to use for matching. Sections 1-4 rely on the choice of method and must be run accordingly for each method chosen. Section 5 compiles matches from all methods and makes the crosswalks.

In order to select the years, set the macros *Year1* and *Year2*. We typically match forward in time so *Year1* is picked as the starting year and *Year2* as the final year. Time difference must also be set between the two years as part of the *timediff* macro. Time difference between two years is calculated as *Year2 – Year1*.

In this code we have the option of using the ABE method to either match using exact names or NYSIIS names that are cleaned for phonetic similarity. To select the type of names to match on set the global *names*.

Section 1: Extract Full Count Census Data by Birthplace

The code uses full count Census data found on the NBER server to make the crosswalks.

The list of birthplaces must be defined in order to extract data by those birthplaces. Because in the original census files birthplaces are coded as longer detailed codes, the floor of birthplace/100 is used to match the list of birthplaces.

Data are extracted and saved by birthplace because when linking large files (e.g. full-count census data), the matching is made more computationally feasible by saving observations by place of birth.

We only keep men in the data. Women cannot be reliably linked over Census decades, particularly between childhood and adulthood because many change their last names after marriage.

Section 2: Standardize variables and clean raw names

The Stata command *abeclean* cleans raw names and generates NYSIIS standardized names. This allows us to match on exact or NYSIIS names in section 3.

Section 3: ABE method using exact names or NYSIIS standardized names

At this step of the matching process there is the option to either match on exact names or match on NYSIIS standardized name. This option is selected by the user in the *Set-Up* section of the code. The ABE algorithm with exact names and NYSIIS standardized names follows the same process.

First the code produces the standard matches resulting from ABE. In the standard version individuals are required to be unique by name (either exact name or NYSIIS standardized name) within each year of birth. The Stata command abematch.ado produces a file containing all successful matches from the standard ABE method. Any variables included in the "keep_A" and "keep_B" will be included in the output file, with the suffix "_A" or "_B", where _A variables come from Year1 and _B come from Year2.

After having produced the standard ABE matches, the code identifies those matches that are unique within ±2 years (the conservative ABE matches). As an output, this section saves the standard and conservative matches made from the ABE algorithm.

Section 4: HISTID Crosswalks for Each Method

This section of the code creates *histid* crosswalks for each method. The *histid* variable is used to create the crosswalks because, unlike the *serial* and *pernum* identifying variables, *histid* remains consistent across different versions of the same Census data. Consistency of *histid* accounts for versioning and allows data to be merged into crosswalks from IPUMS. The crosswalks contain no identifying information.

In Section 4.1. *histid* is merged into matched data from the clean files.

In Section 4.2. files for each birthplace are appended to make a *histid* crosswalk for each method.

Section 5: Creating Master Crosswalks

Individual *histid* crosswalks made using different methods are combined in this part of the code and flags are created for matches made using different methods. The flag *link_abe_NYSIIS_standard* refers to matches made using ABE on NYSIIS names (standard version) and *link_abe_NYSIIS_conservative* refers to matches made using ABE on NYSIIS names but restricting to conservative matches. Similarly, *link_abe_exact_standard* flags matches made using ABE on exact names (standard) and *link_abe_exact_conservative* flags those made using ABE on exact names but conservative.

At this point, the anonymous *histid* to *histid* crosswalk has been made and this crosswalk can be used to merge in publicly available Census data from the IPUMS website. For merging code see *merge_in_variables.do* and *merge_in_variables_readme.*

Last updated: 4/7/20

Please contact ranabr@stanford.edu (Ran Abramitzky), lboustan@princeton.edu (Leah Boustan), and/or myerar@princeton.edu (Myera Rashid) with any questions or feedback about this code.