**How to implement the EM matching method (R)**

A note on external packages: to use these codes you will need to download the following commands from ssc: “nysiis”, “jarowinkler”, “parallel”. Running the following lines in Stata would do the trick: “ssc install nysiis”.

A note on linking methods: this document describes how to implement the EM method using our codes. For a more detailed presentation of this and other algorithms, check: **Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, & Santiago Perez (2019). “Automated Linking of Historical Data”.** (<https://ranabr.people.stanford.edu/sites/g/files/sbiybj5391/f/linking_may2019.pdf>).

**Preparing the data:**

The goal is to link dataset A to dataset B. The linking is based on the following variables:

first\_name

last\_name

Year\_of\_Birth

Place\_of\_Birth

Each observation in each dataset should be identified by an ID variable. We call the ID corresponding to dataset A (ID\_A) and the ID corresponding to dataset B (ID\_B).

**Running the codes (5 steps)**

The codes should be run in the following order:

1. Cleaning and standardizing raw data (clean\_raw\_data.do)

In this initial step, datasets A and B are cleaned and standardized to create their “ready to link” versions.

This step is performed in Stata.

You must set a couple of things before running this file:

* The directory where you have the command “abeclean.ado”.
* The directories where you have your files and the names of these files.
* The directories where you want to save the ready-to-link versions.
* The names you use for your variables: first name, last name, place of birth, and age.
* Within the code, you must also create your id variable. This will depend on how you have created this variable (an example is provided in the code).

2. Create blocks (a\_create\_blocks\_generic.do)

This do-file takes the two datasets that we want to link together and divides them into smaller datasets based on “blocking” variables. We block on: state of birth, first letter of first name and first letter of last name. The output are all the blocks with valid candidates for a match, and also a file that includes all of them together (Data\_Blocks.dta).

This step is performed in Stata.

Here, you must set:

* The directories where you store the ready-to-link datasets, and the names of these datasets.
* The time difference between dataset A and B. (the “timediff” macro).
* The directory where you want to store the blocks.

3. Calculate Jaro-Winkler (JW) distances (b\_jaro\_winkler\_distances\_generic.R)

Takes each of the blocks and computes JW distances and age distances. At this step, I also restrict the set of possible comparisons to observations that differ in at most 5 years in terms of reported year of birth.

This step is performed in R.

Before running this file, you must have the following two directories and set their path in the R file:

* EMblocks: where you stored the blocks in the last step.
* EMdistances: where you want to store the files with candidates for a match and their string and age distances.

4. Expectation-Maximization Algorithm (c\_em\_algorithm\_generic.R)

This step takes as inputs the dataset with all of the distances created in last step and implements the EM algorithm to compute the parameters of the matching problem. It has two outputs: the estimated parameters (EM\_Estimates\_parameters.csv), and the probability of being a match given distances (EM\_Estimates\_probabilities.csv). This last file is used in the next step. Notice that the last files also provides the number of observations for each combination of distances.

This step is performed in R.

Before running this file, you must have the following 2 directories and set their path in the R file:

* EMdistances: where you stored the files with candidates for a match and their string and age distances.
* EMmatches: where you want to store the parameters and probabilities.

5. Apply decision rules (d\_decision\_rules\_generic.do)

We take the estimates from step 4, combined with the data from step 3 to create different datasets based on the choice of the hyperparameters of the linking problem. There are two parameters that we choose: (1) the minimum linking score, and (2) the maximum probability for the second best candidate.

This do-file creates different linked datasets for each combination of parameters.

This step is performed in Stata

Before running this file, you must set the following two directories and set their path in the R file:

* EMdistances: where you stored the files with candidates for a match and their string and age distances.
* EMmatches: where you stored the parameters and probabilities. Matches that come out from the decision rules will be stored here.
* The parameters for the decision rules.

**Which of the linked datasets should be used?**

See Abramitzky, Mill, and Pérez (2019). “Linking Individuals Across Historical Sources: a Fully Automated Approach”. Can be found at this website: <https://ranabr.people.stanford.edu/matching-codes>