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

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

Moreover you must set an id variable. Let’s call it “id”.

**Running the codes**

Before running any of these codes, make sure you have the command to calculate Jaro-Winkler distances in Stata (jarowinkler.ado). Try: ssc install jarowinkler.

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.

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 these variable (an example is provided in the code).

2. Create blocks, calculate distances, EM algorithm, and decision rules (stata\_em\_matching\_example.do)

All these steps are carried out in stata\_em\_matching\_example.do.

Before running this do file you must:

* Set the directory where you stored the ado files you have downloaded (local MatchingDoFiles).
* Set directories where you can find ready-to-link data.
* Set directory to store matches.
* Set directory to store blocks of candidates with JW distances.
* Set the names of your ready-to-link files.
* Set parameters for decision rules (you can try different combinations; see example).
* Set time difference between files (local timediff).

\* Short explanation of “stata\_em\_matching\_example.do”.

1. This file has two main parts.

* In the first one, the command “pairing.ado” is used to create blocks of candidates, and calculates distances. The input of this file are the ready-to-link datasets, and the do file “em\_filter.do”. This last file is the one that calculates the JW distances using the command “jarowinkler.ado” (you must download it before running these files). The output of this file is every block with their respective candidates, and a file that includes all of them (all\_blocks.dta).
* In the second one, the command “matching\_em.do” is used to perform the EM algorithm and apply decision rules. The input is the “all\_blocks.dta” file. You may have different outputs. In the example, we show how to get the estimated coefficients, the estimated probabilities, and the number of observations within each combination of distances, as well as the matches from these procedure.

2. The command “matching\_em.do” can also be used to only apply decision rules. That means, you don’t always need to run the EM algorithm and the decision rules together. In the third section of the example do file we provide an example.

3. You can also choose your own initial guess for the matching probability (for more details, see Abramitzky, Mill, and Pérez (2019)). In the fourth section of the example do file we provide an example.

4. You can set other distributions using the option distributions(). In the example, we use a categorical distribution for each distance.