Matching Methods in Replication Tables
July 21, 2019

Ran Abramitzky, Leah Boustan, and Katherine Eriksson

I. ABE matching methods with exact or NYSIIS standardized names

1. Standard ABE Match with NYSIIS and exact names
   This method allows individuals to iteratively match up to two years in either direction on age but requires an exact match on standardized (NYSIIS) names, and birth state. The "exact" version of this method does not standardize names, apart from fixing common nicknames.
2. Conservative ABE Match with NYSIIS and exact names
   This method restricts the sample so that all individuals are unique by standardize (NYSIIS) or exact names within a five year age band (plus or minus two years) in each census year.
3. Conservative ABE Match with NYSIIS or exact names and exact age
   This match replicates 2 but adds the restriction that matches must be on exact age instead of allowing them to be up to two years off in either direction.

For a more detailed description of these procedures see Abramitzky, Boustan, Eriksson, Feigenbaum, and Pérez (2019, p. 8).

II. ABE matching methods with Jaro-Winkler adjustment

These matching methods initially block by state (or country) of birth, race, and the first letters of both the first and last name. Within a block, we calculate Jaro-Winkler scores for each potential pair. We then use a Jaro-Winkler string distance of 0.1 or less as our cutoff for a match. We also impose certain restrictions on uniqueness within each sample and the number of years by which an individual can misreport age. For example, to be "unique" within a 5 year age band (plus or minus two), there must be no other individual in the same year and block with a Jaro- Winkler distance of 0.1 or less. For a more detailed description of this process see Abramitzky, Boustan, Eriksson, Feigenbaum, and Pérez (2019, p. 11).

5. ABE-JW with exact age matches and uniqueness by exact age.
6. ABE-JW allowing matches up to two years apart in age and with uniqueness within a five-year band (plus or minus two years of age).
7. ABE-JW allowing matches up to two years apart in age, with uniqueness within a five-year band and only keeping matches with the same reported age.

III. EM Algorithm

A detailed discussion of using the Expectation Maximization (EM) algorithm to link individuals across historical sources is included in Abramitzky, Mill and Perez (2019). Choosing which matched individuals to use in analysis depends on two considerations; 1) the match should have a high probability of being true and 2) the second-best match for a given individual should be unlikely to be true. These goals are represented in the choice of parameters p and l. The parameter p determines the minimum probability needed for a pair to be deemed a match, while parameter l is the maximum value we are willing to accept for the probability of the second best match. Higher p and lower l imply lower matching rates and lower false match rates. In our replication results, we present two combinations of parameters according to how conservative we want to be. Since the definition of conservative may differ across papers, each summary table in the replication summaries as well as their corresponding replication tables include the parameter values chosen.

References

Abramitzky, Mill and Perez (2019). Linking Individuals Across Historical Sources: a Fully Automated Approach. This document can be found at: https://ranabr.people.stanford.edu/sites/g/files/sbiybj5391/f/matching_historicalmethod_march27.pdf

Abramitzky, Boustan, Eriksson, Feigenbaum, and Pérez (2019). Automated Linking of Historical Data. This document can be found at: https://ranabr.people.stanford.edu/sites/g/files/sbiybj5391/f/linking_may2019.pdf