ABE/Ferrie matching approach

The ABE/Ferrie matching approach is a method of linking records from one dataset to records in another dataset using first name, last name, and year of birth. The goal is to identify the same set of individuals in both data sets while limiting Type I (mistakenly identifying two different individuals as a match) and Type II (failing to find matched pairs that exist between the data sources) errors. This iterative matching strategy was developed by Ferrie (1996) and adapted and scaled for the computer by Abramitzky, Boustan and Eriksson (2012, 2014, 2017). The Stata codes to implement this matching method are available here.

Standard matching process

The basic steps of the matching process to link one set of records (data set A) to another set of records (data set B) are as follows:

1. First clean names to remove any non-alphabetic characters and account for common mis-spellings and nicknames (e.g. so that Ben and Benjamin would be considered the same name). At this step, any sample restrictions can be imposed.[1] For instance, linking samples are typically restricted to men only because women often change their last names at marriage.
2. We then restrict our attention to only people who are unique by first and last name, birth year, and place of birth (either state or country) in data set A. We do so because, for non-unique cases, it is impossible to determine which of the records should be linked to potential matches in data set B.
3. For each record in data set A, we start by looking for records in data set B that match on first name, last name, place of birth, and exact birth year. At this point there are three possibilities:
    a. If we find a *unique* match, we stop and consider this pair of observations a match.
    b. If we find multiple potential matches in data set B with the same year of birth, we discard this observation (it is impossible to tell which potential match is correct).
    c. If we do not find any match by exact year of birth, we search for matches within ± 1 year of reported birth year, and if this is unsuccessful we look for matches within ±2 years. We only accept unique matches. If none of these attempts produces a unique match, the observation is discarded.

Standard robustness checks

The steps described above represent the basic structure of the ABE/Ferrie matching method. Because historical records are inevitably imperfect, we developed several adaptations of this method to account for different data problems. We recommend testing the robustness of matched samples by using these alternative versions.

1. Requiring matches on exact year of birth. In the standard process we allow matched pairs to differ by up to 2 years in reported year of birth. Alternatively, you can require that matched pairs have the exact same reported year of birth to minimize the chance of false

[1] Note that if you impose an age restriction prior to matching it is necessary to expand the age range by two years on either side (e.g. if you are interested in 20 to 30 year old men, keep in 28 to 32 year old men). The correct age restriction can then be imposed after matches have been found.

positives. However, this will result in a smaller matched sample, and will increase the number of Type II errors.

2. Requiring names to be unique within a 5-year band. In historical records, reported age is an imperfect measure of true year of birth (due to mis-reporting, rounding, and timing of census enumeration). In the standard matching process, we require records to be unique by name and exact year of birth. However due to errors in reported age, checking uniqueness by only exact year of birth can result in false matches. For instance, imagine that two men named James Alexander were born in 1892, but data set A incorrectly reports that one of these men was born in 1890. Both men appear to be unique by name and exact year of birth in data set A, and would be used in the standard matching process. Since in reality these men were born in the same year, they are non-unique and should not be considered in the matching process (it would be impossible to tell when James Alexander is the correct match). By only matching records that are unique by first and last name within a 5-year band around reported year of birth (± 2 years) there is lower risk of false matches due to non-uniqueness. However, since this is a more restrictive uniqueness requirement matched sample sizes will be lower.

3. Using NYSIIS standardized names. Another concern with historical records is misspelling and mistranscription of names. This is risk can be exacerbated when focusing on immigrants with foreign names that census enumerators may not be familiar with. One way of accounting for this is to use the NYSIIS standardized names, rather than exact names, in the matching procedure. The NYSIIS phonetic algorithm standardizes names based on their pronunciation so that names can be matched even if there are minor spelling differences[2].

4. Jaro-Winkler adjustment. Jaro-Winkler string distance gives a measure of the similarity of two strings. An alternative to using NYSIIS standardized names is to compute the Jaro-Winkler string distance between the first and and last names of all potential matches within each data set. The matching process can then decide on the string distance below which any two records are considered as a match. We describe this method in more detail here.

Other matching options

The basic framework of the ABE/Ferrie matching strategy can be altered to fit a variety of contexts. This list describes some other ways the ABE/Ferrie matching method can be altered.

1. Choice of matching variables. For simplicity we described the approach using first name, last name, year of birth, and place of birth to find matches. However, any time-invariant characteristic (such as middle name or mother's name in some contexts) can be used as a matching variable. Note that matching on some time-invariant characteristics (like parent's place of birth or year of immigration to the US) is not necessarily recommended as these are often missing and can be easily misremembered.

[2] The NYSIIS standardization procedure is described in more details here https://en.wikipedia.org/wiki/New_York_State_Identification_and_Intelligence_System and can be implemented with the stata command *nysiis*.

2. Direction of matching. In the standard matching process we describe how to match data set A “forward” to data set B. This refers to the fact that we start by looking at all unique men in data set A, and then search within data set B for potential matches. The direction of matching (A to B vs. B to A) will depend on the context. For instance, when linking between a full-count data set and a smaller data sample, it is best practice to link from the full data set to the sample so that all records considered in the matching process are unique within the full-count data set. Another option is to create a linked sample by matching both “forwards” and “backwards” between the two data sets, and then keep only matched pairs that were identified using both matching directions.