Stanford | Center on Global Poverty and Development

Linking Individuals Across Historical Sources: a Fully Automated Approach

Ran Abramitzky

Roy Mill

Santiago Perez

February, 2018

Working Paper No. 1031

Linking Individuals Across Historical Sources: a Fully Automated Approach*

Ran Abramitzky[a], Roy Mill[b], and Santiago Pérez[c]

[a]Stanford University and NBER
[b]At-Bay Inc.
[c]UC Davis

February 4, 2018

Abstract

Linking individuals across historical datasets relies on information such as name and age that is both non-unique and prone to enumeration and transcription errors. These errors make it impossible to find the correct match with certainty. We suggest a fully automated method for linking historical datasets that enables researchers to create samples that minimize type I (false positives) and type II (false negatives) errors. The first step of the method uses the Expectation-Maximization (EM) algorithm, a standard tool in statistics, to compute the probability that each two observations correspond to the same individual. The second step uses these estimated probabilities to determine which records to use in the analysis. We provide codes to implement this method.

1 Introduction

Linking individuals across datasets offers rich possibilities for economic history research.[1] However, because historical data often lack identifiers such as a Social Security Number, linking individuals relies on personal information such as names and reported ages that is prone to enumeration and digitization errors. These errors make it impossible to find the correct match with certainty. Furthermore, multiple individuals with identical names and reported ages introduce the problem of non-unique matches. Economic historians have developed useful ways to link individuals across historical datasets in the presence of such issues (for example, Atack, Bateman, and Gregson 1992, Ferrie 1996, Abramitzky, Boustan, and Eriksson 2012 and Feigenbaum 2016a; Massey 2017 and Bailey et al. 2017 compare various matching algorithms).

A record matching method should aim to trade-off three goals. First, make as few false matches as possible (minimize type I errors). Second, make as many true matches as possible (minimize

*The first draft of this paper was part of Roy Mill's dissertation completed at Stanford in June 2013. We have benefited from conversations with Jaime Arellano-Bover, Leah Boustan, Raj Chetty, Katherine Eriksson, James Feigenbaum, Tom Zohar and participants in the UC Berkeley complete count census workshop.

[1]Recent examples include Abramitzky, Boustan, and Eriksson 2012, 2013, 2014, 2016; Aizer et al. 2016; Bleakley and Ferrie 2013, 2016; Collins and Wanamaker 2014, 2015, 2017; Eli, Salisbury, and Shertzer 2016; Eriksson 2015; Feigenbaum 2016b, 2017; Ferrie 1997; Fouka 2016; Hornbeck and Naidu 2014; Kosack and Ward 2014; Long 2006; Long and Ferrie 2013; Mill and Stein 2016; Modalsli 2017; Parman 2015; Pérez 2017; Salisbury 2014.

type II errors). Third, for given levels of type I and type II errors, create linked samples that resemble the population of interest as closely as possible. Different research projects may have different implications for compromising on each of these three goals.

We suggest a fully automated method for linking historical datasets that enables researchers to create samples at the frontier of these three goals. The method has two main steps. In the first step, described in sections 3 and 4, we combine distances in reported names and ages between each two potential records into a single score, roughly corresponding to the probability that both records belong to the same individual. We estimate these probabilities using the Expectation-Maximization (EM) algorithm, a standard technique in the statistical literature (Dempster, Laird, and Rubin 1977; Winkler 1989). In the second step, described in section 5, we suggest a number of decision rules that use these estimated probabilities to determine which records to use in the analysis.

This new method for historical record linking helps address concerns about false positives. Moreover, the method is flexible in that it can accommodate different researchers' preferences with respect to the tradeoff between match quality and sample size. We provide the codes that implement the method on the following website:
`https://people.stanford.edu/ranabr/matching-codes`.

2 The matching problem

Imagine you are a researcher who wants to link people from the 1900 to the 1910 census. Imagine that one observation in 1900 is "Ran Abramitzky" who is reported being 10 years old. When you look up this record in 1910, you are looking for a "Ran Abramitzky" who is reported to be a 20 year old. However, when you search the 1910 census, you find three potential matches. One is a "Ran Abramitzky" who is reported to be a 21 year old. One is a "Ran Abramtziky" who is reported to be a 20 year old. And one is a "Ran Abramitzky" who is reported to be a 20 year old.

How would you know which one is the true match? It may be tempting to choose the exact match (third record). However, the other two may as well be the right one given that enumerators can easily make spelling errors and people may not report their exact age but rather round it up or down. An alternative is to declare this record as an impossible to match and drop it from the analysis, but this will result in a smaller sample size.

This problem of record linkage in the presence of errors in identifying information was already discussed almost 50 years ago in statistics (Fellegi and Sunter 1969). Much of this paper simply translates the insights from the statistics literature to the problem of historical record linking.

More formally, we have two datasets containing the description of two populations: A and B. Each pair of individuals from A and B are either a (true) match $M = \{(a,b); a = b, a \in A, b \in B\}$, or a (true) non-match $U = \{(a,b); a \neq b, a \in A, b \in B\}$. Denote the identifying information available on individuals a and b in the datasets as $\alpha(a)$ and $\alpha(b)$. This information can include variables such as names, birthplace, and age.

The researcher needs to come up with a procedure that, based on the information in $\alpha(a)$ and $\alpha(b)$, classifies two records to be either *matched* (M) or *unmatched* (U). There are three goals that need to be taken into account:

1. *Make as few false matches as possible:* This corresponds to minimizing type I errors (minimizing false positives). In other words, we want the least number of cases where the potential match is a false match but we deem it as matched.

2. *Make as many true matches as possible:* This corresponds to minimizing type II errors (mini-

mizing false negatives). In other words, we want the least number of cases where the potential match is a true match but we deem it as unmatched.

3. *Create a sample that is as representative as possible:* For given levels of type I and type II errors, we want the linked sample to resemble the population from which we draw matches as much as possible.

The first two goals describe a standard type 1 versus type 2 error trade-off, and are the ones emphasized in the Fellegi and Sunter 1969 framework. The third goal is an additional challenge that is faced by researchers in the social sciences who are interested in creating linked samples.

3 Selecting identifying and blocking variables and measuring string distances

Before turning into calculating probabilities that each two records are a true match (section 4) and choosing a match to be used in the analysis (section 5), there are three decisions that the researcher has to make. This section discusses these three decisions in turn.

3.1 Selecting identifying variables

The first decision is to choose which identifying variables to use in the matching procedure. The "Ran Abramitzky" example used name and age as identifying variables, but historical datasets often contain other potentially identifying information such as gender, occupation, race, place of birth and place of residence.

The selection of identifying variables will affect all three goals of the match. As we use more variables, we are better able to distinguish between otherwise equally-likely matches. For example, adding age to the list of identifying variables we are potentially able to distinguish between two different Ran Abramitzkys. If we use county of residence, we can distinguish between two Ran Abramitzkys who have the same age. While adding variables to the list of identifying variables may increase the match rates and decrease false match rates, it may also introduce non-representativeness. For instance, a variable like county of residence appears in all censuses and can significantly increase match rates and even help us identify the true individual. However, using such a variable would result in excluding those who switched their county of residence from the analysis. This exclusion will be an issue in a study on geographical mobility, but will not be an issue in a study of fertility among residents who stay in Indiana. Similarly, using occupation for matching will bias any analysis of occupational mobility, but may not be an issue when studying outcomes unrelated to occupations.

The decision of whether to use a variable as an identifying variable thus depends on the research question at hand. In most economics applications, using outcome variables such as occupation or place of residence may be problematic. We suggest following standard practice in economic history and only use predetermined individual level characteristics in the matching procedure. Usually, this restriction reduces the matching variables to names, age and place of birth, which will be the focus of the rest of the paper.[2]

[2]Another variable that could potentially be used in linking is race. However, using this variable could be problematic if individuals selectively report a different race in different historical sources, a pattern documented in Mill and Stein 2016 and Nix and Qian 2015.

3.2 Blocking

The second decision has to do with reducing the computational requirements. In principle, we might want to compare every individual in dataset A to every individual in dataset B. In practice, this is currently not possible computationally unless the size of datasets A and B is very small. The reason is that we would need to perform $n_A \times n_B$ comparisons, where n_A and n_B are the sizes of datasets A and B, respectively. For example, if you need to match 100 records in dataset A to 100 records in dataset B, you will need to make 100*100 = 10,000 comparisons and assign 10,000 probabilities. In a census of millions of people, this can be computationally impractical.

The solution to this computational issue is to only compare individuals who agree on certain *blocking* variables. Ideal blocking variables are those for which mistakes are very unlikely. For instance, if individuals rarely misreport their state of birth, we would be unlikely to miss any true matches by not comparing individuals who declared different states of birth. Further reductions in computational time can be obtained by blocking on gender, or the first letter of the last name. Nevertheless, even though finer blocking results in a lower number of comparisons, blocking is not an innocuous process because it rules out any potential matches across blocks. For instance, if we block by the first letter of the first name, we rule out the name Emmanuel from ever matching to the name Immanuel.

Similar to the choice of identifying variables, the decision on which variables to block on depends on the research question. For example, it will not make sense to block on race in a study of racial passing. Current applications of this method (Mill and Stein 2016; Pérez 2017) restrict the set of comparisons to individuals who are: (1) born on the same state, (2) have the same first letter in first and last names and, (3) have an age difference no larger than five years in absolute value.

3.3 Measuring string distances

The third decision is how to map differences in name spellings into a numerical distance. There is more than one way to compare two strings to each other. One straightforward option is to use an indicator of whether the names are exactly the same. In our example, 1910 "Ran Abramitzky" will have a distance of 0 and 1910 "Ran Abramtziky" will have a distance of 1 from 1900 "Ran Abramitzky". Another option is to use a phonetic algorithm such as NYSIIS instead of the exact name. When using a phonetic algorithm, words that have a similar pronunciation are assigned the same phonetic code. These phonetic codes are designed to overcome name spelling discrepancies that stem from the translation of a heard name to a written name.[3] A third option is to use a continuous string distance measure. When discrepancies in names stem mainly from hearing a name to writing it down, then using phonetic codes such as NYSIIS is a reasonable solution. When the discrepancies come from the exact spelling or digitization of the handwritten record, then string distances can produce better results. Phonetic code match can be used in addition to string distances.

There are many string distance measures available in the literature. We use the Jaro-Winkler string distance (Jaro 1989; Winkler 2006) since it is specifically designed for the comparison of names and was developed in the context of record linking. It calculates a function of the number of matching characters and required transpositions between the two compared strings (names). It gives a higher weight to discrepancies in the first part of the string, where errors are less likely to be made. The original measure is a measure of agreement spanning between 0 (no common characters)

[3]Recent economic history papers use the NYSIIS algorithm. Other examples of phonetic algorithms include Soundex (Odell and Russell 1918) and Metaphone (Philips 1990). Some phonetic algorithms are better suited for dealing with languages other than English. For example, the *Spanish Metaphone* algorithm is designed to match Spanish names (Mosquera, Lloret, and Moreda 2012).

and 1 (exact string match). Since we want to treat all discrepancies in identifying variables as distances, we actually calculate 1 minus the Jaro-Winkler distance as originally defined, thereby having 0 as the distance between two exact names and 1 as the distance between two strings with no common characters.

In the Ran Abramitzky example, "Abramtziky" will be coded as a different name than "Abramitzky" using the NYIIS algorithm, but the Jaro-Winkler distance between these two names will be very low (0.02). At the same time, there are examples in which names have the same NYIIS code but a high Jaro-Winkler distance.[4]

One advantage of string distances is that they are more continuous in nature and can get a wide range of values, unlike a zero/one comparison of exact names or phonetic codes. The wider range of values allows the researcher to conduct sensitivity checks.

4 Assigning a probability that each two records are a true match

After calculating name and other distances such as distances in reported age, we want to combine them into a single distance metric. A natural meaningful measure is the probability that a record pair is a true match. Several ways to estimate this probability have been suggested in non-historical settings (see Winkler 2006 for a rich survey of literature on the subject). In historical settings, Ruggles 2011 and Feigenbaum 2016a estimate these probabilities using a training sample of manually classified records. We suggest an alternative method that does not rely on a training sample, which has the advantage of making the matching easier to replicate by other researchers. The method has been used for record linkage in non-historical contexts and is an application of the Expectation-Maximization (EM) algorithm.[5] This section describes how to apply the EM algorithm to the problem of matching historical records.

To gain intuition about the method, imagine that there are 10 Ran Abramitzkys in the 1900 census, and 10 Ran Abramitzkys in the 1910 census. Each Ran Abramitzky in 1900 is aged from 1 to 10 year old. Our goal is to link these two datasets using information on reported ages, but the challenge is that age is potentially misreported in the 1910 census. For example, somebody who is reported to be 11 in 1900 is reported to be 20 in 1910 instead of 21. This misreporting implies that the age distance will sometimes be greater than zero when comparing two records that belong to the same Ran Abramitzky. Each Ran Abramitzky in 1900 has 10 potential matches in 1910, so we would like to assign a probability that each of these 10 potential matches is the true one. There are 10 Ran Abramitzkys, so there are 100 such probabilities to assign.

To illustrate this example, we simulate 100 age distances. We assume that 10 of these distances correspond to a comparison of true matches, while 90 of them correspond to a comparison of true non-matches. The distances that correspond to true matches are drawn from a normal distribution with mean 0 and standard deviation of 1. The distances that correspond to true non-matches are drawn from a normal distribution with mean 5 and a standard deviation of 1. Panel (a) of figure 1 shows the distribution of observed age distances in this example, if we knew what are true matches and what are non-matches. There are 100 such distances drawn in this graph, each represented as a circle. These age distances come from two different "populations": "matches" (that is, the observations belong to the same individual, corresponding to the 10 circles drawn in red) and "non-matches" (that is, the observations do not belong to the same individual, corresponding to the 90

[4]For instance, "James Tennes" and "James Thomas" have the same NYSIIS code, but the Jaro-Winkler distance between "Tennes" and "Thomas" is 0.4.

[5]The general EM algorithm was described in Dempster, Laird, and Rubin 1977. The specific use of the EM algorithm for record linkage problems was developed by Winkler 1989. For a Bayesian approach to record linkage problems see Larsen 2005.

circles drawn in blue). However, the challenge is that in reality we do not know whether each distance belongs to a comparison of true matches (red) or to a comparison of non-matches (blue). Instead, our actual data look like panel (b) in figure 1. The goal is to use these data to estimate the likelihood that each distance corresponds to a true match, even though we do now know for sure what records are a true match and what records are a non-match.

The EM algorithm starts from assuming that distances between records follow a particular type of distribution, and allowing two different distributions for matches and non-matches. For instance, one possible assumption is that, with probability p_M, distances are distributed normally with mean μ_M and standard deviation σ_M and, with probability $1 - p_M$, distances are distributed normally with mean μ_U and standard deviation σ_U The procedure then estimates p_M, μ_M, σ_M, μ_U, and σ_U, and uses the parameter estimates to identify two separate clusters –one from which true matches are more likely to come and one from which non-matches are more likely to come. Intuitively, we expect age distances to be on average smaller when comparing the same individual than when comparing different individuals. Panel (c) shows the estimated distributions under the assumption that distances are normally distributed. Given these estimated distributions, it is clear that observations that are closer to zero are going to be predicted to be more likely to belong to the population of true matches. In addition, it is clear given the size of each of the clusters that the fraction of true matches (p_M) is smaller than the fraction of true non-matches ($1 - p_M$).

At the same time, the degree of confidence on each of the links will depend on how informative the identifying information (in this case, reported ages) is. The further apart μ_M is from μ_U, the more confident we will be in distinguishing matches and non-matches. Similarly, when σ_M and σ_U are small (that is, if there is very little noise in the identifying information), then we will have more confidence in distinguishing matches and non-matches (there will be less overlap between the estimated distributions).

Imagine now that you try to link both Ran Abramitzky and Santiago Pérez. This will add to the problem the string distance dimension in addition to the difference in reported age. The intuition remains the same, but clustering will be two-dimensional in this case. Figure 2 shows an example in which records differ both with respect to their reported names (x-axis) and ages (y-axis). In panel (a), each data point is labelled as if we knew which records belong to true matches. Panel (b) is how our actual data look like: observations are not labelled as belonging to a comparison of true matches or as a comparison of true non-matches.

More generally, consider the set of ordered pairs of records $A \times B$ and partition this set to the set of true matches (M), if the records in A and B describe the same person, and the complementing set of true non-matches (U). Suppose that the distance, or the degree of non-agreement, in identifying variable k for pair $i \in A \times B$ is given by γ_{ik}, and the vector of such distance measures for pair i is γ_i. Our goal is to estimate for each pair how likely it is to be a true match given the vector of distances in the identifying variables. A pair with shorter distances should be more likely to be a match relative to a non-match.

The probability that a pair i in $A \times B$ is a true match conditional on the distances in the identifying variables γ_i (in our case, reported names and year of birth) can be inferred from Bayes Rule as:

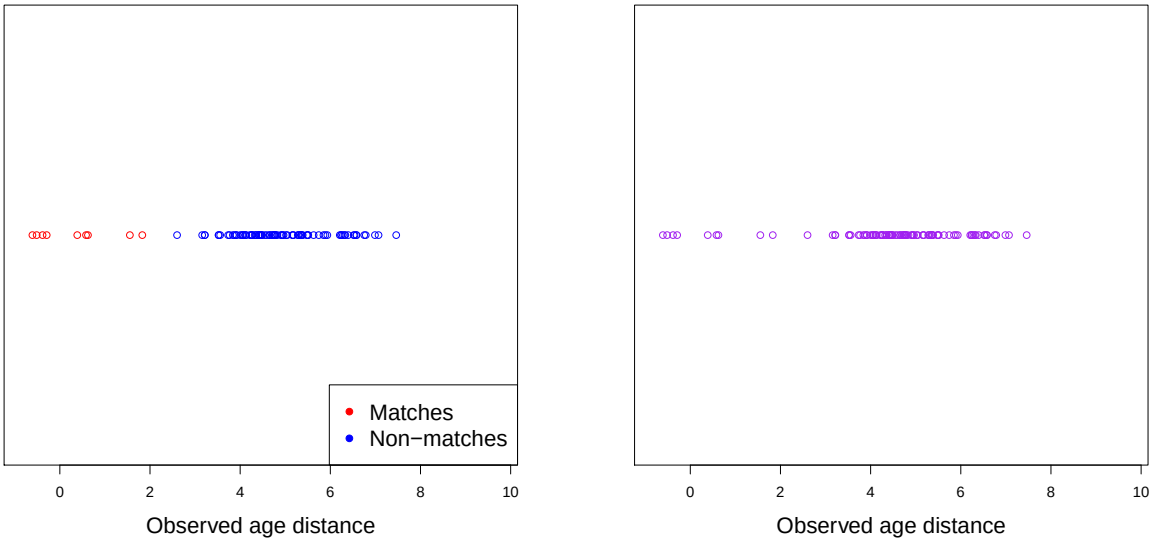
$$\Pr\left[i \in M \,|\gamma_i\right] = \frac{\Pr\left[\gamma_i \cap i \in M\right]}{\Pr\left[\gamma_i\right]} \tag{1}$$

However, we obviously do not really know if pairs are true matches (in M) or non-matches (in U). In other words, pairs are not labeled as being in M or in U. In the data, we observe a sample analogue of $\Pr\left[\gamma_i\right]$ (that is, we observe the empirical distribution of distances across pairs of records, which in our previous example corresponds to panel (b) of figure 1). At the same time, we know that:

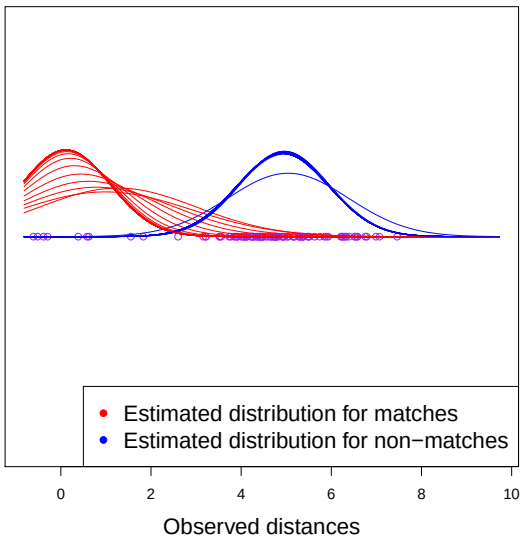
Figure 1: Illustration of the EM algorithm

(a) If true matches were known

(b) Actual data (true matches are unknown)



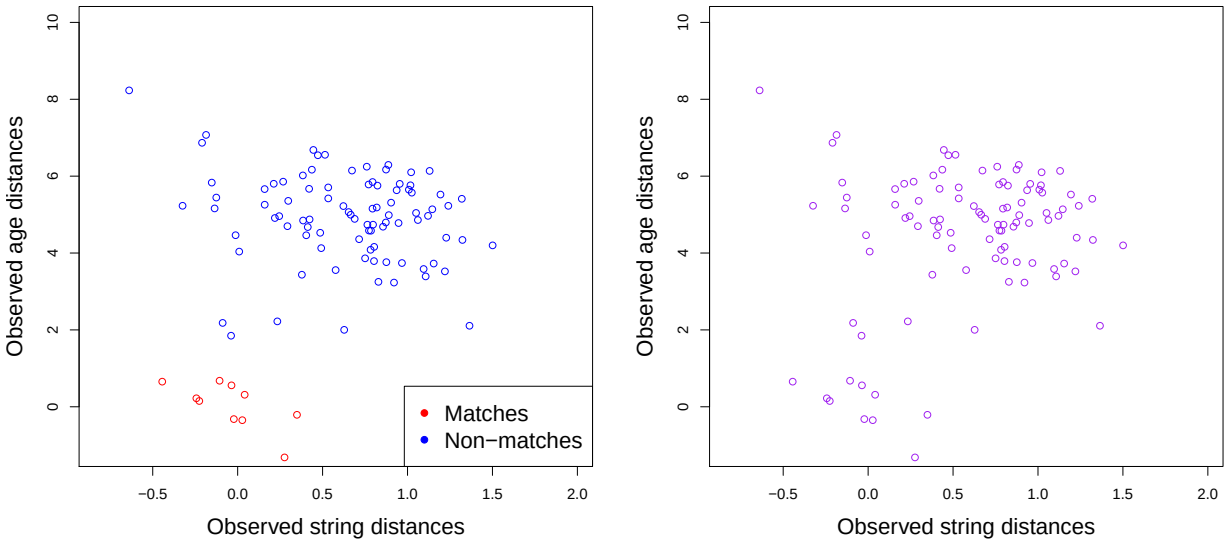
(c) Initial guess and convergence



This figure shows an hypothetical example that illustrates the EM algorithm. Panel (a) shows the situation in which the researchers knows whether the distances correspond to true matches or to true non-matches. Panel (b) shows the actual data, in which true matches are unknown. Panel (c) shows the estimated distributions under the assumption that the distances observed in panel (b) stem from two normal distributions, one corresponding to true matches and one corresponding to true non-matches.

Figure 2: Illustration of the EM algorithm, two-dimensional case

(a) If true matches were known **(b)** Actual data (true matches are unknown)



This figure shows the case in which observations are compared to each other along two dimensions instead: reported ages and names. Panel (a) shows the situation in which the researchers knows whether the distances correspond to true matches or to true non-matches. Panel (b) shows the actual data, in which true matches are unknown.

$$\Pr\left[\gamma_i\right] = \Pr\left(\gamma_i \middle| i \in M\right) \cdot p_M + \Pr\left(\gamma_i \middle| i \in U\right) \cdot (1 - p_M) \tag{2}$$

where p_M is the unconditional probability that a pair is a match.

The method requires that we assume a statistical distribution for $\Pr\left[\gamma_i \middle| i \in M\right]$ and $\Pr\left[\gamma_i \middle| i \in U\right]$. We can then use maximum likelihood to find the parameters of the statistical distribution that maximize the likelihood of observing the observed distances. Once we find these parameters, we can compute:

$$\Pr\left[i \in M \middle| \gamma_i\right] = \frac{\Pr\left[\gamma_i \middle| i \in M\right] \cdot p_M}{\Pr\left(\gamma_i \middle| i \in M\right) \cdot p_M + \Pr\left(\gamma_i \middle| i \in U\right) \cdot (1 - p_M)} \tag{3}$$

That is, the probability that a pair of observations is a match given the observed distances in identifying variables.

If we observed true match status, finding the parameters that maximize the likelihood function would be a straightforward exercise. The reason why we need the EM algorithm to estimate these parameters is because we do not observe true match status, which makes the direct maximization of the likelihood function complicated computationally. The EM algorithm is just a numerical tool that enables us to estimate these parameters without information on true match status.

In particular, the EM algorithm suggests an iterative process to estimate the parameters of the distributions above. It starts by calculating the probability of being a true match (left-hand-side of (1)) given a guess of the distributions of distances conditional on being a match or a non-match (right-hand-side of (1)). Then, based on these probabilities it makes a better guess of the same conditional distribution for another iteration. This process is repeated until the parameters

converge. According to Dempster, Laird, and Rubin 1977 (and specifically in this context according to Winkler 1989) the algorithm reaches a *local* maximum of the likelihood function.

The EM Algorithm

1. Define a distribution family for $\Pr[\gamma_i|\, i \in M], \Pr[\gamma_i|\, i \in U]$. The algorithm will estimate the parameters of the distributions. Denote the vectors of unknown distributional parameters as θ_s where $s \in \{M, U\}$.

   (a) Usually, a conditional independence assumption is invoked: distances in each identifying variable are independent of distances in the other variables (conditional on being a match/non-match). Thus, a distribution can be defined separately for each variable k: $\Pr[\gamma_{ik}|\, i \in M]$ and $\Pr[\gamma_{ik}|\, i \in U]$.

   (b) For binary variables (e.g., state of birth different or not) the distribution is binomial and the parameter of interest will be p_k (the probability to have the same state of birth). For a continuous variable, there could be many families of distributions. If, for example, it is distributed log-normally, then μ_k and σ_k need to be estimated.

2. Guess initial values for parameters of the conditional distributions $\hat{\theta}_s^{(0)}$ and the unconditional probability to be a true match $\hat{p}_M^{(0)}$

3. Loop over steps E and M until convergence:

   (a) *E-step*: Given $\hat{\theta}_s^{(t)}$ and $\hat{p}_M^{(t)}$, infer $w_i^{(t)}$ according to Equation (1):[6]

$$w_i^{(t)} \equiv \Pr\left[i \in M \left| \gamma_i, \hat{\theta}^{(t)}, \hat{p}_M^{(t)} \right.\right] = \frac{\Pr\left(\gamma_i \left| \hat{\theta}_M^{(t)} \right.\right) \hat{p}_M^{(t)}}{\Pr\left(\gamma_i \left| \hat{\theta}_M^{(t)} \right.\right) \hat{p}_M^{(t)} + \Pr\left(\gamma_i \left| \hat{\theta}_U^{(t)} \right.\right) \left(1 - \hat{p}_M^{(t)}\right)} \tag{4}$$

   (b) *M-step*: Given $w_i^{(t)}$, infer $\hat{\theta}_s^{(t+1)}$ and $\hat{p}_M^{(t+1)}$ using Maximum Likelihood. The distribution of γ_i – an observable measure – is given by:

$$\Pr[\gamma_i] = \Pr(\gamma_i|\, i \in M) \cdot p_M + \Pr(\gamma_i|\, i \in U) \cdot (1 - p_M) \tag{5}$$

      i. Hypothetically, if the classification of pairs to true matches and nonmatches was known and denoted by $z_i = I\{i \in M\}$ then we could have estimated $\hat{\theta}_M$ and $\hat{\theta}_U$ from the sample subsets of true matches and nonmatches:

$$\log L(\gamma, \mathbf{z}, \theta, p_M) = \sum_{i=1}^{n} [z_i \log p_M \Pr(\gamma_i|\, \theta_M) + (1 - z_i) \log(1 - p_M) \Pr(\gamma_i|\, \theta_U)] \tag{6}$$

      ii. Since the classification z_i is unknown we replace $\bar{w}_i^{(t)}$ instead of z_i in (6). The

[6] This estimation is also referred to as Maximum A-Posteriori (MAP)

maximum likelihood estimates are then:

$$\begin{aligned}
\hat{p}_M^{(t+1)} &= \frac{1}{N}\sum_{i=1}^{N} \bar{w}_i^{(t)} \\
\hat{\theta}_M^{(t+1)} &= \arg\max_{\theta} \left\{ \sum_{i=1}^{N} \bar{w}_i^{(t)} \log \Pr\left[\gamma_i | \, \theta\right] \right\} \\
\hat{\theta}_U^{(t+1)} &= \arg\max_{\theta} \left\{ \sum_{i=1}^{N} \left(1 - \bar{w}_i^{(t)}\right) \log \Pr\left[\gamma_i | \, \theta\right] \right\}
\end{aligned} \tag{7}$$

After obtaining the maximum likelihood estimates, we can then compute, for any given pair i in $A \times B$ an estimate of:

$$\Pr\left[i \in M \, | \gamma_i\right] \tag{8}$$

The maximum likelihood procedure described above requires assuming an statistical distribution for the observed distances. The distribution selected for the birth year distance was multinomial with six possible outcomes, each corresponding to an age difference ranging from 0 to 5 years in absolute value. Name distances, which are spanning the [0,1] range, were grouped in four ranges following Winkler 1988, roughly corresponding to agreement, partial agreement, partial disagreement and disagreement: [0, 0.067], (0.067,0.12], (0.12, 0.25], and (0.25,1]. We then assumed a multinomial distribution of which range a name distance falls into.

5 Choosing records to use in the analysis

Now that we have estimates of the probabilities that each two records are a true match, we can use these probabilities to choose which matches to use in the analysis. There are several ways to choose a match. One option, for example, is to just choose the match that yields the highest probability of being true. One issue with this approach, however, is that the highest probability can be low, for example 30% of being the true match. Even if the match with the highest probability is very likely (say 90% chance of being the true match), another issue is that there could be a second best match with very similar probability to be the true match (say 80%).[7] A better option is thus to only choose matches with high enough probability to be the true match (say 90%), for which the second best match is unlikely to be the true one (say below 15%).This option will also exclude observations that are non-unique, i.e. observations that have the exact same name and age combination.

Formally, this decision rule can be stated in the following way: To be considered a unique match for a record in dataset A, a record in dataset B has to satisfy three conditions. Specifically, the researcher should:

1. choose the match with highest probability of being a true match out of all potential matches for the record in A.

2. choose a match that is true with a sufficiently high probability, i.e. a match with a probability p_1 that satisfies $p_1 \geq p$ for a given $p \in (0, 1]$ chosen by the researcher.

3. choose a match for which the second best match is unlikely, i.e. the match score of the next best match, denoted as p_2, satisfies $p_1/p_2 \geq l$ for a given $l \in [1, \infty)$ chosen by the researcher.

[7] The EM algorithm does not require these probabilities to add up to 1. That is, for a given record $a \in A$, the sum of the probabilities across all potential matches in B will not necessarily add up to one. The reason is that the EM method does not assume that each observation in a has exactly one match in B.

Similarly, to be considered a unique match for a record in dataset B, a record in dataset A has to satisfy these three conditions.[8] Our linked sample is the set of pairs of records (a, b) in $A \times B$ for which: (1) a matches uniquely to b, and (2) b matches uniquely to a.

Depending on the choice of values for p and l, it is possible to generate samples with more or less confidence on the links. Intuitively, higher values of p and l will yield samples with fewer observations but higher average quality of the links. This possibility enables researchers to assess the robustness of their findings to the quality of the links.

There are analogies between these decision rules and existing automated linking methods in economic history, such as Ferrie 1996 and Abramitzky, Boustan, and Eriksson 2012. When a method requires exact match of the names, it essentially requires that the first best match will have a high enough probability. Similarly, when a method requires uniqueness of the names within a five years window, it essentially requires that the second best match will be unlikely. Requiring both exact match of names and uniqueness within a five years window is parallel to requiring both that the first best match has a high probability and that the second best match is unlikely.

One promising direction not discussed in this paper is how to adjust regression coefficients when dealing with imperfectly linked data. While there is a literature in statistics on this topic (see, for instance, Lahiri and Larsen 2005), these methods are unfortunately still not directly applicable to the situations that typically arise in historical linkage problems. For instance, Lahiri and Larsen 2005 assume that all of the observations in one dataset have a potential link in the other, which does not hold when linking historical censuses due to mortality and underenumeration.

6 Intuition of the method and limitations

As described above, the goal of the method is to split the full set of pairs of records into two groups ("clusters"): matches and non-matches. The simplest way of thinking about this grouping problem would be to use *k-means* clustering. In this approach, the data are split into k clusters so as to (1) minimize the within-cluster differences across observations and (2) maximize the between-clusters differences. Intuitively, pairs of records that are closer to each other with respect to their name and age distances should be grouped together in the cluster of "matches", and observations that are further away should be grouped together in the cluster of "non-matches". The EM algorithm instead computes *probabilities* of observations belonging to each of the clusters. The goal of the method is to maximize the overall probability or likelihood of the data, given the assigned clusters.

Ideally, we would like pairs of records that are close to each other in terms of identifying information to belong to the cluster of matches, while observations that are further apart to belong to the cluster of non-matches. However, a limitation of the approach is that there is no guarantee that the parameters that locally maximize the likelihood function will split the sample into matches and non-matches. Given this, one important sanity check is that the estimated match probabilities are indeed *decreasing* in the distance between observations. Formally, we want that:

$$\gamma_i \leq \gamma_j \Longrightarrow \Pr\left[i \in M \,|\gamma_i\right] \geq \Pr\left[j \in M \,|\gamma_j\right] \tag{9}$$

which, based on (1), is equivalent to having monotone (decreasing) likelihood ratio between $\Pr\left(\gamma|\, i \in M\right)$ and $\Pr\left(\gamma|\, i \in U\right)$:

$$\gamma_i \leq \gamma_j \Longrightarrow \frac{\Pr\left(\gamma_i|\, i \in M\right)}{\Pr\left(\gamma_i|\, i \in U\right)} \geq \frac{\Pr\left(\gamma_j|\, j \in M\right)}{\Pr\left(\gamma_j|\, j \in U\right)} \tag{10}$$

[8] We impose this symmetry condition because linking historical censuses is an example of one-to-one linking. Imposing this condition prevents situations in which a record b in B is the best candidate for a record a in A, but the best candidate for b in B is a different record a' in A.

In the case of conditional independent distributions, this will be satisfied by a monotone likelihood ratio in each of the distances. A more sophisticated version of the code could impose this sanity check as further restrictions on the probabilities (rather than just checking ex post that they are satisfied).[9]

One case in which the algorithm typically fails is when the fraction of true matches (p_M) is very small. One fix to this issue is to use what Yancey 2002 calls a "match enriched sample": a sample in which we oversample observations that are ex-ante more likely to be a true match. One adjustment that works well in practice is to restrict the set of comparisons to individuals who match on place of birth, and first letter of the first and last names. This adjustment largely excludes pairs of records who are very unlikely to belong to the same individual. This issue with the EM algorithm is an additional reason why blocking on some identifying variables is useful.

7 Conclusion

Fully-automated methods for linking historical records are transparent and easy to replicate. We suggest a fully automated method that adapts standard techniques from the statistical literature to the problem of historical record linkage. While this method is more computationally expensive than automated methods based on simple name and age comparisons, it enables researchers to create samples at the frontier of minimizing type I and type II errors. A researcher can choose to create a sample with very low rates of false positives (at the cost of more false negatives), a sample with very low rates of false negatives (at the cost of more false positives), or anything in between.

[9] If there are no duplicates in either datasets A and B, the unconditional match probability p_M cannot be higher than $\frac{min(n_a, n_b)}{n_a \times n_b}$. Hence, another restriction on the parameters that should be checked is whether the condition that $p_M \leq \frac{min(n_a, n_b)}{n_a \times n_b}$ holds.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration". In: *American Economic Review* 102.5, pp. 1832–1856.

— (2013). "Have the poor always been less likely to migrate? Evidence from inheritance practices during the Age of Mass Migration". In: *Journal of Development Economics* 102, pp. 2–14.

— (2014). "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration". In: *Journal of Political Economy* 122.3.

— (2016). "Cultural assimilation during the age of mass migration". In: *National Bureau of Economic Research*.

Aizer, Anna et al. (2016). "The long-run impact of cash transfers to poor families". In: *The American Economic Review* 106.4, pp. 935–971.

Atack, Jeremy, Fred Bateman, and Mary Eschelbach Gregson (1992). ""Matchmaker, Matchmaker, Make Me a Match" A General Personal Computer-Based Matching Program for Historical Research". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 25.2, pp. 53–65.

Bailey, Martha et al. (2017). *How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth*. Tech. rep. National Bureau of Economic Research.

Bleakley, Hoyt and Ferrie (2013). "Up from poverty? The 1832 Cherokee Land Lottery and the long-run distribution of wealth". In: *National Bureau of Economic Research*.

— (2016). "Shocking behavior: Random wealth in antebellum Georgia and human capital across generations". In: *The Quarterly Journal of Economics* 131.3, pp. 1455–1495.

Collins, William J and Marianne H Wanamaker (2014). "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data". In: *American Economic Journal: Applied Economics* 6.1, pp. 220–252.

— (2015). "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants". In: *The Journal of Economic History* 75.4, pp. 947–992.

— (2017). "Up from Slavery? African American Intergenerational Economic Mobility Since 1880". In: *National Bureau of Economic Research*.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38.

Eli, Shari, Laura Salisbury, and Allison Shertzer (2016). "Migration responses to conflict: evidence from the border of the American Civil war". In:

Eriksson, Katherine (2015). "Access to Schooling and the Black-White Incarceration Gap in the Early 20th Century US South: Evidence from Rosenwald Schools". In: *National Bureau of Economic Research*.

Feigenbaum, James J. (2016a). "Automated Census Record Linking: A Machine Learning Approach". In: *mimeo*.

— (2016b). "Intergenerational Mobility during the Great Depression". In: *mimeo*.

— (2017). "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940". In: *Economic Journal*.

Fellegi, Ivan P. and Alan B. Sunter (1969). "A Theory for Record Linkage". In: *Journal of the American Statistical Association* 64.328, pp. 1183–1210.

Ferrie (1996). "A New Sample of Males Linked from the Public Use Micro Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 29.4, pp. 141–156.

Ferrie (1997). "The entry into the US labor market of antebellum European immigrants, 1840–1860". In: *Explorations in Economic History* 34.3, pp. 295–330.

Fouka, Vasiliki (2016). "Backlash: The Unintended Effects of Language Prohibition in US Schools after World War I". In: *Stanford Center for International Development Working Paper* 591.

Hornbeck, Richard and Suresh Naidu (2014). "When the levee breaks: black migration and economic development in the American South". In: *The American Economic Review* 104.3, pp. 963–990.

Jaro, Matthew A. (1989). "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406, pp. 414–420.

Kosack, Edward and Zachary Ward (2014). "Who Crossed the Border? Self-Selection of Mexican Migrants in the Early Twentieth Century". In: *The Journal of Economic History* 74.4, pp. 1015–1044.

Lahiri, Partha and Michael D. Larsen (2005). "Regression Analysis with Linked Data". In: *Journal of the American Statistical Association* 100.469, pp. 222–230.

Larsen, Michael D. (2005). "Hierarchical Bayesian Record Linkage Theory".

Long, Jason (2006). "The Socioeconomic Return to Primary Schooling in Victorian England". In: *Journal of Economic History* 66.4, pp. 1026–1053.

Long, Jason and Ferrie (2013). "Intergenerational occupational mobility in Great Britain and the United States since 1850". In: *The American Economic Review* 103.4, pp. 1109–1137.

Massey, Catherine G (2017). "Playing with matches: An assessment of accuracy in linked historical data". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50.3, pp. 129–143.

Mill, Roy and Luke CD Stein (2016). "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America". In: *Working Paper, Stanford University*.

Modalsli, Jørgen (2017). "Intergenerational Mobility in Norway, 1865–2011". In: *The Scandinavian Journal of Economics* 119.1, pp. 34–71.

Mosquera, Alejandro, Elena Lloret, and Paloma Moreda (2012). "Towards facilitating the accessibility of web 2.0 texts through text normalisation". In: *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pp. 9–14.

Nix, Emily and Nancy Qian (2015). "The Fluidity of Race:"Passing" in the United States, 1880-1940". In: *National Bureau of Economic Research*.

Odell, M and R Russell (1918). "The soundex coding system". In: *US Patents* 1261167.

Parman, John (2015). "Childhood health and sibling outcomes: Nurture Reinforcing nature during the 1918 influenza pandemic". In: *Explorations in Economic History* 58, pp. 22–43.

Pérez, Santiago (2017). "The (South) American Dream: Mobility and Economic Outcomes of First- and Second-Generation Immigrants in Nineteenth-Century Argentina". In: *The Journal of Economic History* 77.4, pp. 971–1006.

Philips, Lawrence (1990). "Hanging on the metaphone". In: *Computer Language* 7.12 (December).

Ruggles, Steven (2011). "Intergenerational Coresidence and Family Transitions in the United States, 1850–1880". In: *Journal of Marriage and Family* 73.1, pp. 136–148.

Salisbury, Laura (2014). "Selective migration, wages, and occupational mobility in nineteenth century America". In: *Explorations in Economic History* 53, pp. 40–63.

Winkler, W. E. (1988). "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage". In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Vol. 667, p. 671.

— (1989). "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage". In: *Proceedings of the Fifth Annual Census Bureau Research Conference*.

Winkler, W. E. (2006). "Overview of Record Linkage and Current Research Directions". In: *U.S Bureau Statistical Research Division Research Report Series* 2.

Yancey, William E (2002). "Improving EM Algorithm Estimates for Record Linkage Parameters". In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Citeseer.