Assessing Individual-Level Record Linkage between Historical Datasets

Roy Mill

March 2012
Very preliminary

Linking individual-level records between different historical datasets has been on the rise in economic history recently. Researchers can now construct longitudinal datasets of individuals and dynasties across censuses.[1] However, until the middle of the 20th century, individuals were not allotted universal personal identification numbers like Social Security numbers. Identifying individuals relies on personal details that are both non-unique and prone to enumeration and digitization errors.[2] The non-uniqueness and misreporting introduce uncertainty in the process of matching.

When such uncertainty arises, researchers face a tradeoff between erroneously deeming two unrelated records as a match (Type I error) and erroneously neglecting true matches that were slightly misenumerated (Type II error). The solution to the problem of uncertain linkage can be divided to two interrelated parts: First, how to use as much information as possible to assess the likelihood of a record pair to be a match (while keeping selection bias minimal)? Second, how to incorporate the link's uncertainty into the analysis of interest? This document restricts its attention only to the former problem of assessing link reliability. It suggests a few improvements over current methods prevailing in economic history. The latter problem of adjusting the analysis to incorporate the uncertainty in the linkage is outside the scope of this paper and is discussed, for example, in Abramitzky et al. [2011a], Scheuren and Winkler [1993], Lahiri and Larsen [2005].

This paper suggests three interrelated improvements and assesses the benefit of their use: First, it suggests using the Jaro-Winkler (Jaro [1989], Winkler [1990]) method to measure distances between the name searched and names of potential matches. Second, it suggests combining distances in names and birth years to a single statistic, roughly corresponding to a probability to be a true match, using EM algorithm (Dempster et al. [1977], Winkler [1989]). Finally, a decision rule is suggested for deciding which potential matches to take as matches for further analysis.

Table (1) summarizes the problem using two cases of siblings in the 1910 census that are matched to the 1930 census. George and John Sales appeared in the 1910 census as

[1] See Abramitzky et al. [2011b, 2012, forthcoming], Collins and Wanamaker [2011], Ferrie [1996, 1997a,b, 2003a,b], Ferrie and Long [forthcoming, 2007], Long [2005, 2006], Stewart [2006] among others.

[2] Although unique, even Social Security Numbers can be erroneously reported. According to Winkler [2006] about 1-2% of individuals misreport their Social Security numbers in quarterly employment data of the State of California.

TABLE 1: DESCRIPTION OF THE PROBLEM THROUGH A CASE STUDY

Potential matches in 1930 for George and John Sales, aged 12 and 7 in 1910

	Potential match details			NYSIIS codes		Distance in				Match by	
#	First name	Last name	Age	First name	Last name	First names	Last names	Implied birth year	"Prob" of match	iterative method?	race*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1.1	GEORGE	SAYLES	36	GARG	SAL	0	0.04	4	0.52	0	B/Mu
1.2	GEORGE	SEILS	36	GARG	SAL	0	0.20	4	0.46	0	B/Mu
1.3	GEORGE	SCHOOLS	31	GARG	SAL	0	0.39	1	0.42	1	W
1.4	GEORGE	SHIELS	37	GARG	SAL	0	0.24	5	0.40	0	B/Mu
1.5	GEORGE	SCHOOLS	27	GARG	SAL	0	0.39	5	0.38	0	W
2.1	JOHN	SAYLES	26	JAN	SAL	0	0.04	1	0.55	0	B/Mu
2.2	JOHN	SAULS	26	JAN	SAL	0	0.11	1	0.49	0	W
2.3	JOHN	SILLS	31	JAN	SAL	0	0.24	4	0.41	0	B/Mu
2.4	JOHN	SILLS	31	JAN	SAL	0	0.24	4	0.41	0	B/Mu
2.5	JAMES	SELES	27	JAN	SAL	0.47	0.20	0	0.12	1	W

- Rows 1.1-1.5 are potential matches for George Sales. Rows 2.1-2.5 are potential matches for John Sales.
- Columns (5) and (6) show the NYSIIS code of the potential match's names.
NYSIIS codes standardizes names by ignoring minor typographical errors.
- Columns (7) and (8) present (1 - Jaro-Winkler) distances between the names of the original 1910 record and the potential match.
The distance counts non-mutual letters and transpositions of mutual letters.
- Column (10) presents EM-algorithm estimates that the record is a match as a function of name and birth year distances.
- The iterative method matches the record with the same NYSIIS codes of first and last names and the lowest birth year difference, if that lowest-birth-year-difference record is unique.
* Neither matching procedures took race into account. Race is used here as an indicator of possible false matches.

aged 12 and 7, racially classified as "Mulatto" and "Black," respectively, and both born in Virginia. Searching in the 1930 census for individuals born in Virginia, with an implied birth year no farther than five years away and the same (phonetically standardized) names, yielded five candidate matches for each of the brothers. The current methods in economic history either define the record with the closest birth year as a match or drop the two siblings altogether from the matched file. The former method is referred to as the iterative match method and the match indicator is presented in column (11). The methods we propose produce a "probability" of – or a measure of confidence in – a potential match being a true match. This measure is presented in column (10).

Looking at the table it seems that the iterative match did not produce the likeliest match. Moreover, both of them matched to a record of a white person. We did not use race in the matching procedure because racial classification can change over the course of time, but discrepancies in it between linked records can be used as a proxy false match rates.

Even though all the names translate to the same phonetic code (columns (6) and (7)), some of the names appear to be closer than others. Not taking into account these differences and restricting attention to the difference in birth year ignores important information. Apart from using additional information, our method allows the researcher to decide whether the "best" match is also far enough from the "second best" match to be deemed as a match, or to use the probabilities of all potential matches in subsequent statistical analysis that we will not discuss.

We still offer a way to resolve uncertainty for some of the cases and show that a conservative calibration of this rule can lead to substantially lower false matching rates (and

lower overall matching rates).

1 Existing linkage methods in economic history

The convention in economic history has been to follow Ferrie [1996], where a true match is defined there as having:

1. The same standardized (and phonetically coded) first and last name.
2. No more than 10 such name combinations in the population (regardless of age).
3. The difference in the implied birth years is no more than five years.
4. The birth state (or country) is the same.
5. If the person headed a household in the 1850 census, they should head the same household in 1860
6. If more than one potential match satisfy the previous five conditions, the potential match with the minimum age difference was linked. In Ferrie [1996]'s case, this resolved all multiple potential matches (for name combinations that had 10 or less cases overall to begin with).

Other papers diverged from this method in three ways. First, the prerequisite of no more than 10 records with the same name is no longer maintained as indexes with birth year are digitized. Second, information on household members is no longer relied upon as it is creating further selection bias into the matched sample. Third, resolving the ambiguity by taking the record with the lowest difference in birth year is maintained in Abramitzky et al. [2011b, 2012, forthcoming] – and is called the iterative matching procedure there – but is no longer used in Long [2005, 2006], Ferrie and Long [2007], Long [2008], Ferrie and Long [forthcoming], where the observation is dropped when there is more than one potential match within a birth year band of five years.

Before moving on to describe possible improvements to these methods, it is worth mentioning that some papers, like Ruggles [2011], use the recently devised linked census files offered by IPUMS (see Ruggles [2006], IPUMS [2010]). These files show pairs of records linked between the full 1880 census and one of the historical censuses samples. The linked files use a measure of string distance between first and last names and then use a machine-learning algorithm to translate the distances to a likelihood ratio of match-vs.-nonmatch. Potential links that exhibit a high likelihood ratio are then classified as matches.

The classification of the algorithm depends on a small subset of potential pairs that were classified by highly trained clerks as true matches or true nonmatches. The problem with involving human judgment in the decision of whether to link or not is that it is not documented in a way that allows the researcher to assess possible reasons for bias, it is not replicable, and it depends on the skills and incentives of clerks. Thus, a fully automatic algorithm, like the one described in Section (2.2), is preferred.

So far the literature treats the merged dataset in subsequent analysis as if the matched records are certain. We will follow this convention and only produce an alternative method to pick the best match and to decide whether to use it in the analysis or not. Suggestions for improvement on this front are found in Abramitzky et al. [2011a], Lahiri and Larsen [2005], Scheuren and Winkler [1993, 1997].

2 Possible improvements

Even though the current methods work quite well for many cases, they can still be improved. We will first describe a more continuous measure of agreement, or distance, between names. A more continuous measure allows to do to names what existing methods already do to birth year difference: finely tune the decision rule. Then we will describe a way to translate the distance measures to a single measure corresponding to the probability to be a true match. This probability can be used literally in regressions of interest, but such analysis is out of the scope of this document (see Scheuren and Winkler [1993, 1997], Lahiri and Larsen [2005] for methods that use the probabilities to be a true match in subsequent regressions). We will treat the probability estimates as a single statistic measuring overall agreement between two records in every potential match. Based on this measure, a decision rule for whether to consider a potential match as a (true) match is then proposed. We will show how parameters of this decision rule affect matching rates and false-match rates.

2.1 String distances

One major improvement that can be made to the current measurement of name agreement is the use of string distances. Names are currently translated to NYSIIS codes that try to group spellings of similarly sounding names. While the NYSIIS code successfully catches misspellings of names for a very large number of cases, codes are still quite raw and lose a lot of information that can be extracted from a comparison of names.

The Jaro-Winkler distance (Jaro [1989], Winkler [1990]) calculates a function of the number of matching characters and required transpositions between the two compared strings (names). It gives a higher weight to the first part of the string. We calculate 1 minus the distance as it is originally defined, thereby having 0 represent the distance between two exact names and 1 represent the distance between two strings with no common characters.

Examples for *possible* Type I errors – incorrectly classifying a pair of records as links (false matches) – are presented in Table 2a. The table shows selected potential matches where the NYSIIS codes for the first and last name match, but the string distance for either of them is quite high.

Examples for *possible* Type II errors – incorrect classification of links as nonlinks – are presented in Table 2b. These are cases where even though distance between the strings is quite low, NYSIIS codes do not match.

More important than reclassification of matches is the more continuous nature of string distances. Unlike the binary comparison between NYSIIS code, a string distance allows

TABLE 2: CASES WHERE NYSIIS CODE MATCH AND NAME DISTANCES DIVERGE

(a) Potential matches with long name-distance that still match on NYSIIS code

Raw name				Distance		NYSIIS code of names (first 4 chars for first)			
Source		Potential match		First	Last	Source		Potential match	
JOHN	HART	JOHN	HEYWARD	0	0.48	JAN	HAD	JAN	HAD
JOHN	ROOKS	JOHN	REECE	0	0.48	JAN	RAC	JAN	RAC
HARRY	HEYM	HARRY	HAINOE	0	0.48	HARY	HAN	HARY	HAN
JAMES	TENNES	JAMES	THOMAS	0	0.40	JAN	TAN	JAN	TAN
WILLIAM	SYMON	WILLIAM	SHANNON	0	0.39	WALA	SANAN	WALA	SANAN
ANNA	SEALS	ANNE	SCHULZ	0.12	0.38	AN	SAL	AN	SAL
ROBERT	TYNES	ROBERTA	THOMAS	0.03	0.38	RABA	TAN	RABA	TAN
BENJAMIN	ROUSAW	BENJAMIN	REESE	0	0.38	BANJ	RAS	BANJ	RAS
WILLIAM	SLATE	WILLIAM	SCHLETT	0	0.26	WALA	SLAT	WALA	SLAT
JOHN	BUCKLEY	JOHN	BEEKLEY	0	0.26	JAN	BACLY	JAN	BACLY
JOHN	ALLEN	JOHN	ALOUNAH	0	0.26	JAN	ALAN	JAN	ALAN
JOHN	STORD	JOHN	STEWART	0	0.26	JAN	STAD	JAN	STAD
THOMAS	CALEMON	THOMAS	COLENAN	0	0.26	TAN	CALANAN	TAN	CALANAN

(b) Potential matches with short name-distance but no match on NYSIIS code

Raw name				Distance		NYSIIS code of names (first 4 chars for first)			
Source		Potential match		First	Last	Source		Potential match	
HALBERT	CLARK	ALBERT	CLARK	0.05	0	**HALB**	CLARC	**ALBA**	CLARC
HANNA	WILLIAMS	ANNA	WILLIAMS	0.07	0	**HAN**	WALAN	**AN**	WALAN
ANNIS	JOHNSON	ANNIE	JOHNSON	0.08	0	**AN**	JANSAN	**ANY**	JANSAN
ARTHUR	BURRIES	ARTHUR	BURRIS	0	0.03	ARTA	**BARY**	ARTA	**BAR**
ARTHUR	HARRIS	ARTHUR	HARRISS	0	0.03	ARTA	**HAR**	ARTA	**HARA**
ALFRED	HARDEMAN	ALFRED	HARDMAN	0	0.03	ALFR	**HARDANAN**	ALFR	**HARDNAN**
ARTHUR	EMMEY	ARTHUR	EMME	0	0.04	ARTA	**ANY**	ARTA	**AN**
ANDREW	WILLIAMSON	ANDREW	WILLIAMS	0	0.04	ANDR	**WALANSAN**	ANDR	**WALAN**
ALBERT	HILL	ALBERT	HILLY	0	0.04	ALBA	**HAL**	ALBA	**HALY**
ANDREW	WILLINGHAM	ANDREW	WILLIAMS	0	0.09	ANDR	**WALANGAN**	ANDR	**WALAN**
ARTHUR	HAWKS	ARTHUR	HAWES	0	0.09	ARTA	**HAC**	ARTA	**HA**
ARTHUR	JEFFRIES	ARTHUR	JEFFERSON	0	0.10	ARTA	**JAFRY**	ARTA	**JAFARSAN**

sensitivity checks to be performed. As in the case of birth year, where the researcher can decide what band of years around the original record's birth year and how to choose the potential match that is closest to the original birth year, such exercises are possible when using string distances but are impossible using a binary rule.

2.2 Combining distance measures to a probability of a true match

We usually consider distances in more than one identifying variables when we link historical datasets. How can one turn the distances between each of the variables' values to a single metric?

A natural meaningful measure can be 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). A very popular method is an application of the Expectation Maximization (EM) algorithm.[3] This section describes how to apply the EM algorithm to the problem of matching records in historical censuses.

There are two files, A and B, containing records with data on people that may appear in both datasets. There is a group of common identifying variables (e.g. names, birth year, etc.). 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 nonmatch.

The probability that a pair is a match conditional on the distances in the identifying variables can inferred from Bayes Rule:

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

where p_M is the unconditional probability that a pair is a match. The problem is that pairs are not labeled by their true match/nonmatch status: this was the problem in the first place.

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 nonmatch (right-hand-side of (1)). Then, based on these probabilities it makes a better guess of the same conditional distribution for another iteration. The process stops when the guessed probabilities of being a match or not yield estimated parameters that yield the same probabilities. According to Dempster et al. [1977] (and specifically in this context according to Winkler [1989]) the algorithm reaches a local maximum of the likelihood function.

[3]The general EM algorithm was described in Dempster et al. [1977]. The problem of record linkage and an optimal decision rule were formally defined by Fellegi and Sunter [1969]. 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].

2.2.1 The Algorithm

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

   (a) Usually, as in our case, 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: $\Pr[\gamma_{ik} | i \in M], \Pr[\gamma_{ik} | i \in U]$.

   (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):[4]

$$w_i^{(t)} \equiv \Pr\left[i \in M \left| \gamma_i, \hat{\theta}^{(t)}, \hat{p}_M^{(t)} \right.\right] = \frac{\Pr\left(\gamma_i \left| \hat{\theta}_M^{(t)} \right.\right) \hat{p}_M^{(t)}}{\Pr\left(\gamma_i \left| \hat{\theta}_M^{(t)} \right.\right) \hat{p}_M^{(t)} + \Pr\left(\gamma_i \left| \hat{\theta}_U^{(t)} \right.\right) \left(1 - \hat{p}_M^{(t)}\right)} \tag{2}$$

      i. To guarantee convergence, Winkler [1989] suggests using a constraint in the form of $\hat{p}_M^{(t)} \leq p_0$. If $\hat{p}_M^{(t)} = \frac{1}{n}\sum_{i=1}^n w_i^{(t)} > p_0$ then all weights are rescaled by multiplying them by $\bar{w}_i^{(t)} = w_i^{(t)} \cdot \frac{p_0}{\frac{1}{n}\sum_{i=1}^n w_i^{(t)}}$.[5]

   (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 – can be given by:

$$\Pr(\gamma_i | \theta, p_M) = p_M \Pr(\gamma_i | \theta_M) + (1 - p_M) \Pr(\gamma_i | \theta_U) \tag{3}$$

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

$$\log L(\gamma, \mathbf{z}, \theta, p_M) = \sum_{i=1}^n [z_i \log p_M \Pr(\gamma_i | \theta_M) + (1 - z_i) \log(1 - p_M) \Pr(\gamma_i | \theta_U)] \tag{4}$$

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

[5] One natural choice is the number of records searched for divided by the number of returned results. To improve results, p_0 should not be too low. It can also be perturbed to see how sensitive is the process to the choice.

ii. Since the classification z_i is unknown we replace $\bar{w}_i^{(t)}$ instead of z_i in (4). The maximum likelihood estimates are then:

$$\begin{aligned}
\hat{p}_M^{(t+1)} &= \frac{1}{N}\sum_{i=1}^{N}\bar{w}_i^{(t)} \\
\hat{\theta}_M^{(t+1)} &= \arg\max_{\theta}\left\{\sum_{i=1}^{N}\bar{w}_i^{(t)}\log\Pr\left[\gamma_i|\,\theta\right]\right\} \\
\hat{\theta}_U^{(t+1)} &= \arg\max_{\theta}\left\{\sum_{i=1}^{N}\left(1-\bar{w}_i^{(t)}\right)\log\Pr\left[\gamma_i|\,\theta\right]\right\}
\end{aligned} \tag{5}$$

Imposing further restrictions on the probabilities A reasonable requirement from estimated match probabilities is that they are decreasing in the distances: whenever distances are (unambiguously) shorter the probability of a match is higher. Formally,

$$\gamma_i \leq \gamma_j \Longrightarrow \Pr\left[i \in M\,|\gamma_i\right] \geq \Pr\left[j \in M\,|\gamma_j\right] \tag{6}$$

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

$$\gamma_i \leq \gamma_j \Longrightarrow \frac{\Pr\left(\gamma_i|\, i \in M\right)}{\Pr\left(\gamma_i|\, i \in U\right)} \geq \frac{\Pr\left(\gamma_j|\, j \in M\right)}{\Pr\left(\gamma_j|\, j \in U\right)} \tag{7}$$

In the case of conditional independent distributions, this will be satisfied by a monotone likelihood ratio in each of the distances. This requirement translates to constraints for the maximum likelihood estimation of the parameters in the M-step. The constraints depend on the distribution in question. In the results section below we estimated the parameters with no constraints, and luckily the estimated distributions satisfied the decreasing LR requirement.

2.2.2 Estimation of match-probability results

This section uses the same data used for Mill and Stein [2012]. The sample in 1910 consists of all sons classified as "Mulatto" in the 1910 census who had another son in their household with a different racial/color classification.

We have 46,162 individuals in the 1910 census that we want to match to the 1930 census. Some of these individuals have two versions of their names (digitized by two clerks) and therefore we queried 58,798 names. Since we did not have the complete census index we used Ancestry.com's search engine and intentionally set the search to be very expansive. Ancestry.com returned roughly 800,000 hits back. We limited the amount of scraped hits to 100 per person so in our file we have 603,748 pairs of potential matches.

Before showing the results of the EM algorithm, Table 5a reports how many of the potential matches are considered unique by different conventional standards. The benchmark matching to which we will compare will be the one denoted "iterative matching" following Abramitzky et al. [forthcoming]. One can see that the iterative method maximizes the number of names that are matched uniquely. However, the maximization of the number

TABLE 3: MATCHING RATES IN ANCESTRY.COM AND FURTHER MATCHING RULES

Matching rule	Hits per name	Proportion of names with 0 hits	1 hit	2+ hits
Ancestry.com's search	13.91	0.377	0.131	0.492
Ancestry.com's search truncated	10.27	0.377	0.131	0.492
NYSIIS(names) equal, birth year diff ≤ 5	4.98	0.500	0.166	0.335
NYSIIS(names) equal, birth year diff ≤ 2	2.39	0.586	0.172	0.242
NYSIIS(names) equal, birth year diff ≤ 1	1.45	0.651	0.164	0.185
NYSIIS(names) equal, birth year diff = 0	0.52	0.782	0.124	0.094
NYSIIS(names) equal, iterative birth year	0.93	0.500	0.333	0.167

* Ancestry.com's search truncated caps the number of hits for a name at 100

* Iterative birth year matches lowest birth year difference found for a name

of unique matches may also increase the number of erroneously matched pairs. Our decision to choose a record that has the same birth year rather than another potential match where the birth year is only one year away may be erroneous, especially since individuals reported age rather than birth year and a deviation of one year is not unlikely.

After restricting potential matches to be born in the same state as the searched record (and further restrictions by Ancestry.com's expansive search), we ran the EM algorithm on distances in three identifying variables: first name, last name and birth year. Distances in last and first names were measured using the Jaro-Winkler distance measure.[6]

The distribution selected for the birth year distance was naturally multinomial. Name distances, which are spanning the $[0, 1]$ range, were grouped to 9 ranges: $\{0\}$, $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$, $(0.4, 0.5]$, $(0.5, 0.6]$, $(0.6, 0.7]$, and $(0.9, 1]$ (no distances between 0.7 and 0.8 were found) and a multinomial distribution of which range a name distance falls into was assumed. The EM algorithm then estimated the set of probabilities for each of the three variables' values.

Table 4 presents how the distances are distributed in the data independent of their match status in column (3). Columns (4) and (5) present the distributions conditional on being a match or a nonmatch as given by the EM algorithm. Column (6) presents the likelihood ratio and it is evident that it is decreasing in the distance for each of the three variables. Figure 1 displays the distribution of the estimated probabilities. Bear in mind that the whole estimation process and the distribution of distances are a function of Ancestry.com's search engine's output.

2.2.3 A matching rule based on the probability measures

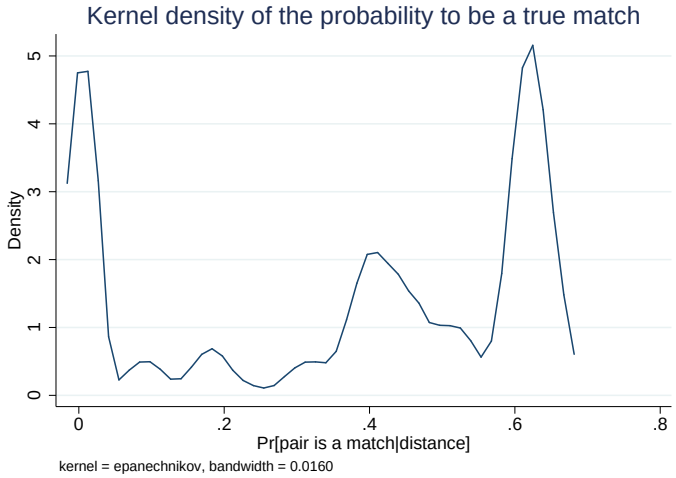
Out of the 46,162 people searched in the 1930 census, 32.62% have no potential match returned by Ancestry.com. Another 14.15% had more than one match where the best two matches have the same probability to be a true match. This is mostly due to them having the same distance measures in all variables. In the first case there is objectively nothing to do and in the second a decision rule based solely on the probability measure cannot yield a

[6]Again, the distance measure was actually 1 minus the Jaro-Winkler distance measure and therefore equal to 0 if identical, to 1 if no common characters exist, and $\gamma_{ik} \in (0, 1)$ for partial agreement.

Table 4: Unconditional and estimated conditional distributions of distances

Distances		Unconditional	Estimated probability conditional on being		Likelihood ratio M/U
Variable	Distance	proportion	Match	Nonmatch	$\frac{\Pr[\gamma\|M]}{\Pr[\gamma\|U]}$
(1)	(2)	(3)	(4)	(5)	(6)
Last name	0	0.511	0.632	0.441	1.431
	(0,.1]	0.067	0.066	0.068	0.971
	(.1,.2]	0.115	0.097	0.126	0.771
	(.2,.3]	0.124	0.093	0.143	0.649
	(.3,.4]	0.128	0.091	0.149	0.614
	(.4,.5]	0.040	0.021	0.050	0.424
	(.5,.6]	0.009	0.0004	0.014	0.027
	(.6,.7]	0.000	1.0e-06	0.0005	0.002
	(.9,1]	0.006	0.0000	0.009	0.000
First name	0	0.593	0.864	0.437	1.980
	(0,.1]	0.023	0.032	0.017	1.866
	(.1,.2]	0.015	0.018	0.012	1.488
	(.2,.3]	0.026	0.031	0.023	1.338
	(.3,.4]	0.036	0.028	0.041	0.688
	(.4,.5]	0.065	0.023	0.090	0.260
	(.5,.6]	0.093	0.003	0.145	0.021
	(.6,.7]	0.012	0.0004	0.018	0.020
	(.9,1]	0.138	6.3e-07	0.218	0.000
Birth year	0	0.100	0.111	0.093	1.194
	1	0.181	0.190	0.176	1.081
	2	0.188	0.190	0.186	1.020
	3	0.181	0.176	0.183	0.962
	4	0.177	0.171	0.181	0.945
	5	0.174	0.161	0.181	0.893

Figure 1: Distribution of estimated probabilities to be a match $\Pr[i \in M \,|\gamma_i]$



unique match. 53.23% of the sample remains to be matched according to a decision rule.

A natural rule for selecting the unique match takes the potential match with the highest probability if that probability is above a certain threshold, and if it is bigger than the next best potential match's probability by some measure. In other words, sort each searched person's potential matches in 1930 according to their estimated probability to be a match from highest to lowest. Then define pair i of searched person j (1 being best pair, I being worst pair) to be a match if $i = 1$, $\hat{p}_1 > \underline{p}$ and $\frac{\hat{p}_1}{\hat{p}_2} > \underline{l}$. If there is no second potential match, define $\frac{p_1}{p_2} = \infty$.

The first restriction – a minimum probability to be deemed a match – is set to minimize false matches even if there is only one potential match. The second restriction – the likelihood ratio between the first and second best matches – is set to resolve ambiguity when more than one match has a good chance to be a match, especially when names are common.

The parameters $\underline{p}$ and $\underline{l}$ will determine both matching rates and false match rates: higher $\underline{p}$ and $\underline{l}$ imply lower matching rates and lower false match rates. Table 5a shows the proportion of individuals that are matched by the matching rule, as a function of the parameters. Matching rates are increasing to the right and bottom of the table, as the cutoffs are relaxed. For comparison, the iterative matching method of Abramitzky et al. [forthcoming] applied to this data yields a matching rate of 0.333 (see Table 3). This high matching rate corresponds to the bottom right part of Table 5a.

But increasing the match rate does not come without cost. False matches are more likely to occur when we are less certain of the top match or when there are more than one potential match with reasonable probabilities to be a match. We can use a variable in the data that is indicative of someone being the same person, like the race variable in Table 1. Since race wasn't available for all potential matches (due to the scraping technology), I use county of residence in 1910 and in 1930. We don't use race and location as identifying variables in the matching process because they are likely to bias our sample, but we can use them as measures of false matching rates. Even though people can change their racial identification or move to another counties, many tend to stay, and we expect a better matching technique (that does not use these variables for the matching decisions) to yield higher rates of agreement.

Table 5b presents the proportion of pairs that live in the same county for pairs that are deemed matched with the parameters in question. One can see the same pattern as in Table 5a: agreement decreases (false rates increase) when we move from left to right and from top to bottom. The translation from the proportion of pairs in different counties to the proportion of false matches depends on the proportion of migrants and the probability that a random match (correlated by name, age and birth state) will be in the same state. However, the pattern of increasing false match rates commensurate to overall match rates is clear.

The proportion of same-county pairs out of all matched pairs by the iterative method stands at 0.16 and corresponds to the bottom right end of the table as well. The proposed decision rule can be calibrated to produce a desired combination of sample size and possible false match rate. A decrease in the matching rate from 0.333 to 0.2 can correspond

to an increase in the same county pairs from 0.16 to 0.2. A more conservative matching technique can yield only 0.1 matching rate, but 0.281 in same-county proportion.

TABLE 5: MATCH AND ERROR RATES AS FUNCTIONS OF DECISION RULE'S PARAMETERS

(a) Matching rates, out of 46,162 searched individuals

	For those with best Pr[match | distances] above...								
	0.55	0.5	0.45	0.4	0.35	0.3	0.25	0.2	0.15
$\underline{l}$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
∞	0.061	0.077	0.088	0.103	0.113	0.117	0.119	0.120	0.122
2.5	0.073	0.092	0.105	0.123	0.136	0.142	0.144	0.146	0.150
2	0.074	0.094	0.107	0.126	0.140	0.146	0.149	0.151	0.156
1.75	0.077	0.098	0.111	0.130	0.145	0.152	0.155	0.157	0.162
1.5	0.092	0.114	0.128	0.147	0.164	0.171	0.174	0.176	0.182
1.4	0.103	0.126	0.141	0.161	0.177	0.184	0.188	0.191	0.196
1.3	0.116	0.142	0.157	0.178	0.195	0.203	0.206	0.210	0.215
1.2	0.135	0.166	0.183	0.205	0.223	0.231	0.235	0.238	0.244
1.1	0.158	0.197	0.220	0.246	0.266	0.274	0.278	0.282	0.288
1.05	0.205	0.253	0.283	0.315	0.336	0.345	0.349	0.354	0.360

$\underline{l}$ is the cutoff likelihood ratio between the best and second-best potential match's probabilities

(b) Proportion of *matched* pairs that are in same county in 1910 and 1930

	For those with best Pr[match | distances] above...								
	0.55	0.5	0.45	0.4	0.35	0.3	0.25	0.2	0.15
$\underline{l}$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
∞	0.296	0.270	0.249	0.219	0.202	0.195	0.193	0.191	0.189
2.5	0.288	0.265	0.245	0.216	0.198	0.191	0.188	0.186	0.182
2	0.288	0.265	0.245	0.215	0.197	0.190	0.187	0.184	0.180
1.75	0.290	0.266	0.246	0.217	0.197	0.190	0.186	0.184	0.179
1.5	0.288	0.267	0.249	0.221	0.202	0.195	0.192	0.189	0.185
1.4	0.281	0.260	0.244	0.219	0.201	0.194	0.191	0.189	0.185
1.3	0.272	0.251	0.237	0.214	0.198	0.191	0.188	0.186	0.182
1.2	0.262	0.241	0.228	0.208	0.194	0.188	0.185	0.183	0.180
1.1	0.240	0.220	0.206	0.189	0.177	0.172	0.170	0.168	0.165
1.05	0.215	0.198	0.186	0.171	0.162	0.159	0.157	0.155	0.153

$\underline{l}$ is the cutoff likelihood ratio between the best and second-best potential match's probabilities

References

Ran Abramitzky, B. Douglas Bernheim, Aprajit Mahajan, and Isabelle Sin. A simple procedure for improving the use of imperfectly linked data. 2011a.

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. Leaving the nest: Lessons from 19th century norway on migration and the childhood environment. October 2011b.

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. April 2012.

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, forthcoming.

William Collins and Marianne H. Wanamaker. Selection and economic gains in the great migration of african americans: New evidence with linked census data. 2011.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, December 1969.

Joseph P. Ferrie. 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. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29 (4):141–156, 1996.

Joseph P. Ferrie. The entry into the U.S. labor market of antebellum European immigrants, 1840–1860. *Explorations in Economic History*, 34:295–330, 1997a.

Joseph P. Ferrie. Immigrants and natives: comparative economic performance in the united states, 1850- 1860 and 1965- 1980. *Research in Labor Economics*, 16:319–341, 1997b.

Joseph P. Ferrie. The rich and the dead: Socioeconomic status and mortality in the U.S., 1850–1860. In Dora L. Costa, editor, *Health and Labor Force Participation over the Life Cycle: Evidence from the Past (National Bureau of Economic Research Conference Report)*. University Of Chicago Press, 2003a.

Joseph P. Ferrie. Internal migration. In Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright, editors, *Historical Statistics of the U.S., Millennial Edition*. Cambridge University Press, 2003b.

Joseph P. Ferrie and Jason Long. The path to convergence: Intergenerational occupational mobility in britain and the u.s. in three eras. *Economic Journal*, 117(519):C61–C71, March 2007.

Joseph P. Ferrie and Jason Long. Intergenerational occupational mobility in britain and the us since 1850. *American Economic Review*, forthcoming.

IPUMS. Ipums linked representative samples, 1850-1930 final data release, June 2010.

Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406): 414–420, June 1989.

Partha Lahiri and Michael D. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, Mar 2005.

Michael D. Larsen. Hierarchical bayesian record linkage theory. August 2005.

Jason Long. Rural-urban migration and socioeconomic mobility in Victorian Britain. *Journal of Economic History*, 65(1):1–35, March 2005.

Jason Long. The socioeconomic return to primary schooling in Victorian England. *Journal of Economic History*, 66(4):1026–1053, December 2006.

Jason Long. Social mobility within and across generations in britain since 1851. 2008.

Roy Mill and Luke C.D. Stein. Disentangling skin-color discrimination from family background differences: African-americans in the early 20th century. February 2012.

Steven Ruggles. Linking historical censuses: A new approach. *History and Computing*, 14: 213–224, 2006.

Steven Ruggles. Intergenerational coresidence and family transitions in the united states, 1850–1880. *Journal of Marriage and Family*, 73(1):136–148, 2011.

Fritz Scheuren and William E. Winkler. Regression analysis of data files that are computer matched. *Survey Methodology*, 19:39–58, 1993.

Fritz Scheuren and William E. Winkler. Regression analysis of data files that are computer matched, part ii. *Survey Methodology*, 23:157–165, 1997.

James I. Stewart. Migration to the agricultural frontier and wealth accumulation, 1860–1870. *Explorations in Economic History*, 43(4):547–577, 2006.

William E. Winkler. Near automatic weight computation in the fellegi-sunter model of record linkage. In *Proceedings of the Fifth Annual Census Bureau Research Conference*, 1989.

William E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1990.

William E. Winkler. Overview of record linkage and current research directions. *U.S Bureau Statistical Research Division Research Report Series*, 2, 2006.