IPUMS
Working Papers

A New Strategy for Linking Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel

Jonas Helgertz†
University of Minnesota, Lund University

Joseph R. Price
Brigham Young University

Jacob Wellington
University of Minnesota

Kelly Thompson
University of Minnesota

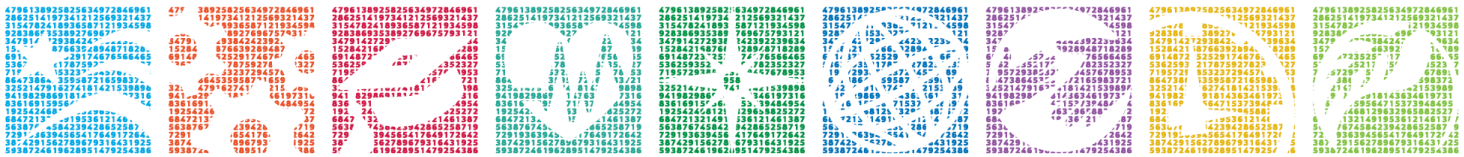
Steven Ruggles
University of Minnesota

Catherine R. Fitch
University of Minnesota

September 2020

Working Paper No. 2020-03
DOI: https://doi.org/10.18128/IPUMS2020-03

†Address correspondence to Jonas Helgertz: helgertz@umn.edu. This research was funded by the National Institute on Aging grant R01AG057679. Financial support from the Minnesota Population Center is also acknowledged, through core funding (P2C HD041023) from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (NICHD). Comments and suggestions from Matt Sobek, Dave Hacker, Evan Roberts, John Robert Warren, Matt Nelson, Leah Boustan, Ran Abramitzky and James Feigenbaum are gratefully acknowledged.



Abstract

This paper presents a new probabilistic method of record linkage, developed using the U.S. full count censuses of 1900 and 1910 but applicable to a range of different sources of historical records. The method was designed to exploit a more comprehensive set of individual and contextual characteristics present in historical census data, aiming to obtain a machine learning algorithm that better distinguishes between multiple potential matches. Our results demonstrate that the method achieves a match rate that is twice as high other currently popular methods in the literature while at the same time also achieving greater accuracy. In addition, the method only performs negligibly worse than other algorithms in resembling the target population.

Introduction

There is a broad set of social research questions that benefit from linking individuals across datasets over time. This includes research related to demographic behavior, intergenerational mobility and how conditions during childhood affects later life outcomes. In the United States, considerable efforts have been made to digitize vast amounts of individual level data, with the publicly available full-count decennial censuses from 1850to 1940 being the most notable example. This data opens the possibility for life-course studies across multiple birth cohorts, but the lack of a stable personal identifier (such as a Social Security number) in historical sources makes it challenging to link individuals across records.

During the past 25 years, multiple efforts have been undertaken to combine historical individual records to create longitudinal data, allowing for an improved understanding of the past. Along with digitization of historical records and improved computational capabilities, the past decades have seen an acceleration of the development of computerized, automated linking methods. In this paper, we present a new probabilistic method of record linkage, developed for linking individuals across historical records using the U.S. full count censuses of 1900 and 1910. The method proposed has implications for how data should be processed and exploited in order to maximize precision and linkage rates. More specifically, our method exploits the contextual nature of census data (individuals within households, households within neighborhoods), both when generating input data for the machine learning algorithm (training data), as well as when designing characteristics subsequently used by the linking algorithm in order to better distinguish between multiple potential matches. In doing so, not only are we able to achieve a match rate that is twice as high as the methods currently used in the literature, but we also achieve a higher degree of precision and obtain a linked sample which does equally well in resembling the target population. While the method outlined in the paper has been developed for linking individuals across decennial U.S. censuses, its underlying principles can also help improve the linking of other datasets in which individuals are listed along with their family members or neighbors.

The linking process described in this paper incorporates additional household and neighborhood contextual features, yielding computer-generated matches closer to those a human genealogist would produce and – equally importantly - without decreasing performance for individuals lacking this data. This is achieved through the implementation of a two-phase linking approach which first links individuals, then links the remaining individuals among linked households. This can provide greater recall within a population while maximizing precision/accuracy.

Background

Efforts to create longitudinal individual data for historical populations, frequently exploiting numerous different types of source material, are by no means novel. Examples include longitudinal datasets from Sweden (Scanian Economic Demographic Database, POPLINK), Canada (BALSAC), the Netherlands (Historical Sample of the Netherlands) and China (China Multi-Generational Panel Datasets), allowing for the study of demographic and socioeconomic outcomes from a life-course perspective over time periods, and in some cases going as far back in time as the 17th century. The databases typically cover limited geographical areas, only allowing for the examination of individuals while they reside in the area in question. Despite their limitations, the databases represent extremely valuable resources, being the result of large teams of research assistants manually linking individuals across sources, also facilitated by the rather localized area covered by the database.

Record linkage of historical U.S. censuses began as early as Malin (1935) who linked farm operators in Kansas across thirteen censuses taken between 1860 and 1935. All early census record linkage was limited to smaller geographical areas such as Trempeleau county, Wisconsin (Curti 1959), Newburyport, Massachusetts (Thernstrom 1964), Atlanta, Georgia (Hopkins 1968), Boston, Massachusetts (Knights 1971) and Kingston, New York (Blumin 1976). The overarching linking methodology in these studies was to track a population of men over time by manually searching microfilmed census listings. Since their source material only covered the location where the sample population was first observed, the study population became restricted to individuals who did not migrate, with nontrivial consequences regarding representability of the resulting sample (Ruggles et al. 2018).

While constant human supervision over the linking process is an advantage, there are also many disadvantages. The true characteristics of the individual in the record are often altered by reporting errors from the respondent, errors introduced by the enumerator, and transcription errors in converting the hand-written text into machine-readable format. Given these sources of errors, the decision of whether individuals in two records are the same person often involves weighing a complicated set of factors. As a consequence, it is difficult for manual linkers to maintain consistent criteria for determining matches. It is difficult to fully document the decision process, making replication impossible. Beyond the concerns about validity and reliability, manual record linkage is time consuming and expensive, making it infeasible for large populations that extend beyond local areas.

While earlier attempts to track individuals over time using U.S. data relied on source material with limited opportunities to sort, restrict, and search among records, the first release of IPUMS data in the early 1990s represented a watershed moment. The digitization of historical census records opened new opportunities to

examine and follow substantially larger populations over time, catalyzing the transition to computerized approaches for processing the data. With the availability of larger population sets, one major challenge associated with linking census data became obvious: how could one systematically and most effectively use the available information in the source material to find the records corresponding to the same individual in another source? Theoretically, this is a straightforward task, as we expect individuals to carry certain immutable characteristics over their lifespan. For example, an individual named John Smith, male and born in state *s* in the year *t* will display these same characteristics in the next census. Defining the universe of potential matches would therefore represent a straightforward task, limiting the population to individuals sharing those characteristics. Several factors make this less straightforward, however, including the prevalence of proxy reporting and lack of detail and precision in the data. As a result, nontrivial differences in the spelling of names and in the reported year or state of birth are common. This presents the researcher with a dilemma, since while allowing for a widely defined universe of potential matches increases the probability that the true match will be among the potential matches, it also increases the risk of a Type I error[1] as well as increasing computational requirements. For example, going from restricting potential matches to individuals reporting the same year of birth to instead allowing for birth year reporting error of +/- three years increases the number of potential matches (amount of data that needs to be processed) by 900 percent, holding everything else constant.

The first comprehensive attempt at linking U.S. census data was made by Ferrie (1996), whose approach originated from a set of rules based on which records across two different censuses were considered to be the same person. Ferrie exploited the IPUMS sample for 1850 and an alphabetic name index of the 1860 census that had been constructed for genealogical use. Ferrie coded both the IPUMS data and the 1860 index phonetically and searched the 1860 index for cases that phonetically matched each name in the IPUMS sample. After discarding cases with more than 10 potential matches, Ferrie located each potential match on the microfilm of the census enumeration, and used birth year (within three years), state or country of birth, and the presence of family members to determine which match was correct. If there were two or more perfect matches, the individual with the closest age difference was selected.

At a meeting on historical record linkage held at the University of Montreal in 2003, Ruggles (2006) argued that using information about family members, place of residence, or occupation to disambiguate potential links would introduce selection bias that would be likely to distort estimates of geographic or economic mobility and other life-course transitions. Ferrie had come to a similar conclusion, and he announced that

[1] Type I error refers to incorrectly declared matches, also called false positives.

he had already embarked on a new fully-automated record linkage project using only characteristics that in theory should remain consistent over time: name (for males), age, sex, and birthplace.

Ferrie's fully-automated record linkage became feasible with the advent of the first full-count historical census microdata. In 2003, a collaboration of IPUMS and The Church of Jesus Christ of Latter-day Saints released census microdata covering the entire U.S. population enumerated in 1880, comprising over 50 million records (Roberts et al. 2003). Ferrie (2005) linked the new 1880 full-count database to the IPUMS 1% samples of the 1850, 1860, 1870, 1900, and 1910 censuses. To avoid selection bias, Ferrie considered a limited set of variables. He required an exact match on name (except for very small spelling variations), matching birthplace, and a birth year within three years. All multiple matches were dropped, and no information on family members, place of residence, or other variables was consulted.

Subsequent efforts to link U.S. censuses have virtually all followed Ferrie's lead and avoided the use of variables that could introduce selection biases. IPUMS linked the 1880-full count census to the other IPUMS samples using the same variables as Ferrie, but using a probabilistic machine-learning strategy instead of his deterministic approach. The IPUMS Linked Representative Samples (IPUMS-LRS) used a support vector machine to obtain the predicted probability that two records are a true match, based on a set of characteristics believed to be time-invariant across the life course, as well as consistently being available in both sets of data that were being linked (Goeken et al 2011). The support vector machine used manually linked input data to calibrate the relative importance of each examined characteristic, using characteristics such as name commonality, first and last name similarity score, and age difference. This probabilistic method of record linkage thereby allows for a more flexible linking approach, but also one whose performance ultimately will depend on the accuracy of the input data used by the algorithm to recognize patterns in the data that is consistent with a pair of records referring to the same individual. Under the umbrella of IPUMS-LRS, multiple datasets were released, spanning the time period 1850-1930, with versions of the method also being applied to census data from other countries.

In recent years, automated linkage of U.S. censuses has become increasingly common. Many studies adapted and scaled Ferrie (1996) for use with the full-count census records that made the linking replicable, fully automated, and transparent. The first paper to implement this type of census linking for the entire population was Abramitzky, Boustan, and Eriksson (2012), and a similar approach was used by Abramitzky, Boustan and Eriksson (2014), Collins and Wanamaker (2015), Beach et al. (2016), and Alexander and Ward (2018). In general, these studies combine the use of phonetic classification used in Ferrie (1996) with the rules introduced by Ferrie (2005). Recent work uses statistical algorithms such as expectation maximization to determine which links are correct (Abramitsky, Mill, and Pérez 2018; Pérez

2019). The code needed to implement these approaches is now widely available for researchers to use (censuslinkingproject.org) and has become a very important tool for linking historical records.

Following the example of IPUMS-LRS, other investigators have turned to probabilistic machine-learning approaches. Feigenbaum (2016) adapted a regression model to training data using statistical software and methods that are in the wheelhouse of many social scientists in order to evaluate potential matches in his 1915 Iowa sample, linked to the 1940 Census. His approach made record linking methods that employ machine learning more accessible and more understandable. Other efforts to link historical records using supervised machine learning methods include Bailey et al. (2019), Abramitzky et al. (2020); and Price et al. (2019). The precision of each of these approaches is based on the training data that can be used to teach the model and the machine learning algorithm to discern between true and false links. Bailey (2018) provides a description of a massive effort to create training data using humans to label true and false links between historical records that involve frequent standardized training and double- and triple-entry practices to ensure a high level of quality. Price et al. (2019) uses links created on a public genealogy platform to create training data in which high quality links are created by people doing family history for their relatives and using information that goes beyond the fields contained in the census records.

In the wake of the emergence of several straightforwardly applicable methods of record linkage and the availability of a plethora of digitized individual-level historical data, the focus has shifted towards comparing how existing methods perform. There is considerable debate about the quality of links created through automated methods (Bailey et al. 2020; Abramitzky et al. 2020). Bailey et al. (2020) uses the large training set that they created to evaluate the quality of automated linking methods. They note that 15 to 37 percent of the links created by automated methods are identified as false links by human reviewers. Their evaluation highlights the importance of comparing the predictions of automated methods with the decisions made by humans doing the same task to evaluate possible improvements to automated approaches to link records. Hand linking individuals across records is a much too slow and expensive process to provide the primary approach, but efforts to create training data or validation sets can be used in combination with automated methods to create linked samples with high match rates and high precision.

Our approach

All the early efforts to link historical records—from Malin to the first iteration of Ferrie—used all information available for linking, including the characteristics of household members. These projects generally linked only a small fraction of the population for two main reasons. First, most studies were local and lost track of out-migrants; second, the source data were not machine-readable, and it was impossible to broadly search on many characteristics to find the best possible match. Because only a small percentage of

the population was linked, there was high potential for selection bias, especially bias favoring non-migrants who resided with the same family members across multiple census years. To mitigate these biases, record linkage projects conducted since 2003 have used only time-invariant characteristics, mainly name, birth year, sex, and birthplace.

In the past seven years, IPUMS has released full-count machine-readable data for every surviving U.S. census from 1850 to 1940. The availability of the full-count data opens the potential for a new approach to record linkage. A consistent feature of the U.S. historical censuses, conducted every ten years, is that information on a range of individual level demographic and socioeconomic characteristics was collected, and the information is organized into families and households. By using all the information available—mutable and immutable—we can link a far higher percentage of the population than was previously possible, with far lower levels of false links.

Selection bias remains a concern. People who remain in the same place or remain married to the same person, for example, have more information available to establish links than do those who migrate without kin. All linkage efforts, however, introduce selection biases, and our preliminary analysis suggests that the bias introduced by our approach is comparatively small. Our linked data complements work based on linking individuals using immutable characteristics. The benefits of higher linkage rates and improved precision should be weighed against the concern of potential selection bias, particularly for social and geographic mobility.

We use a range of individual, household, and contextual characteristics in a machine-learning probabilistic record linkage algorithm. We use a two-step approach, where we begin by linking men, obtaining a sample of high confidence and high precision links. We then proceed to a second step, exploiting the household links that are generated in stage one. Here, we use household-links generated in step one to maximally restrict the universe of potential matches, using a modified machine learning algorithm to link household members–men and women–that were not linked in the first stage. Each stage requires a unique set of training data to calibrate its respective machine learning algorithm, and is outlined separately below.

Step I: Linking Men

Training Data

For the generation of the initial set of hand-linked training data, we extracted a sample of 3,000 men from the 1900 census. The training data consists of 50 randomly selected men from each state of birth (50*50=2500), in addition to 50 randomly selected men from 10 different regions of origin outside the United States (50*10=500). While deviating from the standard procedure when generating training data by

not being a random draw from the underlying 1900 population of men, the sample nevertheless largely reflects the full 1900-population in terms of basic demographic characteristics, other than region of birth.

The universe of potential matches extracted from the 1910 full-count census was generated by restricting the linkable population to individuals who were male and born in the same state within +/- three years of the 1900 individual. In addition, the universe was limited to those having at least one identical last name adjusted bigram[2] and a (unstandardized) first and last name Jaro-Winkler score of at least 0.7, respectively. Out of the 3,000 individuals initially selected from the 1900 census, the enforcement of aforementioned blocking criteria[3] results in a population of 2,700 potentially linkable individuals.

We set out to generate training data of the highest quality possible, through systematically relying on the wealth of resources provided by Ancestry.com when evaluating each potential match. Out of the 2,700 individuals extracted from the 1900 census, we were able to confidently link 1,354 individuals, or 50.1 percent to a record among the universe of 1910 census potential matches. The linking rate is somewhat difficult to compare to figures reported in other historical research on the United States due to differences in linking methods, samples and sources, but we remain confident that our share represents a gold standard.

Expanding the Universe of Linking Variables

Our machine learning algorithm shares several fundamental procedural characteristics with those proposed by Feigenbaum (2016). In terms of linking variables, our linking algorithm benefits from a number of those used by Feigenbaum, while also considerably expanding the set of linking variables. Indeed, this reflects one key extension of our linking approach, proposing that data–in particular census data spanning comparatively shorter time periods–can be used more effectively and enhance the algorithm's ability to more accurately distinguish between potential matches as outlined in greater detail below.

An illustrative example of the underlying idea of our algorithm is provided in Figure 1, taking as a point of departure a frequently encountered situation when relying solely on individual level information. In the example, the individual to be linked, Thomas P Arthur, represents someone with a relatively common name, also translating to more than one perfect match on first and last name to individual records in the 1910 census. In a situation like this, remaining time-invariant linking characteristics are unlikely to assist in

[2] The adjustment of the bigrams is through the first letter of the name being its own token. For example, Helgertz is split into "H", "He", "el", "lg", "ge", "er", "rt", "tz". Consequently, net of the Jaro-Winkler score similarity, the last names Helgertz and Wellington would pass this criterion, through "el". Within the linking software developed to implement the algorithm outlined in this article, the use of first and last name adjusted bigrams reduces the baseline number of computations required during the generation of the universe of potential matches by 87 percent on sample datasets, reducing the runtime by about 70 percent.

[3] In addition, an update of the 1910 full-count census resulted in a slight further reduction (n=29) in the number of 1900 individuals, due to an adjusted year of birth variable.

further distinguishing between the potential matches. As a result, all of the first four potential matches emerge as equally plausible candidates, with the linking algorithm thereby failing to identify one unique match. If anything, a human doing this task might choose the bottom row since the birth year matches exactly and there is no conflicting information for the middle initial.

- Figure 1 here

One key idea of this paper is that additional and readily available information can and should be used in order to arrive at a better calibrated machine learning algorithm. Figure 2 elaborates on the contents of Figure 1 in order to illustrate the usefulness of broadening the set of information used by the algorithm. In the example below, information on the name of the father of the 1900 individual as well as of the 1910 potential match is added, allowing for straightforwardly determining what by all accounts appears to be the correct match. In addition, it then becomes clear that the P and F for the middle initial are written in a very similar way and one of them is likely a transcription error. Indeed, while in many situations it remains very difficult to conclusively determine one potential match as the correct one, the idea underlying our algorithm is that being able to rely on additional information increases the degree of precision in doing so.

–Figure 2 here

Parents and spouse characteristics

Much of the historical census population is embedded in households and families that often persist across multiple censuses. This is particularly true for children under the age of 7 and married adults who are often with at least some of the same family members ten years later.

For every record, categorical variables indicate whether the individual's mother, father, or spouse was present in both the household of the 1900 individual and the 1910 potential match. When this is the case, the variable additionally indicates whether there is substantial mismatching information on key characteristics, suggesting that the 1900 and 1910 records are referring to a different parent or spouse. More specifically, the variable indicates the presence of (substantial) mismatch on year or place of birth, as well as whether (for the spouse) there are unrealistic values on the marriage duration variable or (for the parents) whether the parent's relationship to the target individual changes, i.e. from biological to step-parent or vice versa. For observations where both the 1900 individual and the 1910 census potential match has a non-missing observation on the name of the family member in question, we additionally calculate the Jaro-Winkler name similarity score.

Another potentially relevant piece of information available for all individuals in the 1900 and 1910 U.S. censuses is provided by the birthplace of both the individual's mother and father. This is operationalized as

two separate indicators showing whether this matches across the censuses for the individual's mother and father, respectively.

Other household members

Albeit less straightforward to operationalize in a manageable way, similarly useful information in accurately identifying matching records may be provided by other household members, related or otherwise. Due to individuals frequently residing in large households, we opted to operationalize the information by distinguishing between related (i.e. siblings, grandparents) and unrelated (lodgers, etc.) household members. Within each category, we calculate the Jaro-Winkler score for each member from the 1900-household and compare it to every age (+/-5 years) and sex-appropriate individual belonging to that same household member type in the potential match's 1910 household. Thus, the Jaro-Winkler name similarity of a female relative to the 1900 index individual named "Anna" and born 1884 is calculated for all female relatives of the 1910 potential match that were born between 1879 and 1889. This ascertains that comparisons are performed between the relevant "category" of people and, furthermore, that an index individual's uncle, born in the mid-1800s will not be compared to a potential link's sister, born 1895, not only belonging to a different sex but also widely differing in age. Subsequent to performing all relevant comparisons, an indicator variable that is used by the algorithm is generated, signaling whether there is at least one relative of the 1900 individual whose JW name similarity score is greater than or equal to 0.9 when compared to a qualifying 1910 potential match's related household member. Consequently, this indicates a high likelihood that there is (at least) one common relative who resides in the same household as the 1900 individual and the potential match in 1910. An analogously generated indicator variable measures the presence of unrelated household members that are present in both records.

Residential characteristics

The censuses also contain information on the individual's place of residence and on potentially relevant neighborhood characteristics. While not available for every household, many individuals in both censuses live in households where the street name is reported. Conditional on the 1900 index individual and the 1910 potential match individual living in the same state and county, we calculate the degree of street name similarity[4], again through the Jaro-Winkler string comparison score. The rationale underlying this as a linking variable pertains to its use in *confirming* rather than *rejecting* a potential match. More specifically, since migration was not uncommon, a low street name similarity score fails to provide any evidence *against*

[4] Street names are cleaned, removing universal components, including "street", "avenue", "road" and "alley"

a potential match. On the other hand, however, a high score should provide strong evidence in favor of a match.

A second indicator is represented by calculating the share of common neighbors between the 1900 index individual and the 1910 potential match individual. Again, its main expected use is in confirming rather than rejecting potential matches, which is why we carefully design the variable only to assist with the former. We begin by extracting the ten nearest preceding and ten following household heads' last names for the 1900 sample individual, conditional on the neighboring household residing in the same county and state as the sample individual. We follow the same procedure for the potential matching individual from the 1910 census. Thus, for an individual with at least ten households listed before and after on the census form, respectively, and residing in the same county and state, the individual's twenty closest residing neighbors are obtained. Conditional on the 1900 individual and the 1910 potential match residing in the same state and county in both censuses, through Jaro-Winkler scores, we proceed to compare each of the 1900 individuals' neighbors to every neighbor household's last names of the potential match from 1910. Treating JW-scores above 0.95 as evidence of the presence of a neighbor, a 1900 sample individual will have between zero and twenty common neighbors. Since the vast majority of observations pertain to individuals with either *zero* or *several* common neighbors, we operationalize this information as a dichotomous variable, indicating whether one or more neighboring household is identical across the census records.

Additional linking variables

When linking a sample of the population from one record to the full population in another record, the risk of declaring false positives will increase. An illustrative example is provided by a situation where the sample that we are trying to link contains an individual *John Stevenson*, born in state *s* and in year *y*. Born in the same state and year is a second John Stevenson who was not included in the random sample of the population that we chose to link. If only the second John Stevenson survived through the linking period, resulting in the linking algorithm almost certainly declaring a positive match with the namesake that was included in the linking sample. It is easy to see how this problem is exacerbated with sample data, since if we were linking full populations, both John Stevensons would likely be linked to the only surviving one. Through appropriate post-processing of the linked data, dropping duplicate links, both would be discarded from the data. In an attempt to quantify the extent of this problem for each 1900 sample individual and allowing for the algorithm to account for this, we calculate the number of individuals sharing the 1900 sample individual's (standardized) first and last name in the 1900 census, conditional on being born in the same country/state and within +/- three years. The underlying logic is to provide the algorithm with a

quantification of the likelihood that a link could be declared to another individual with the exact same name, given the aforementioned blocking criteria.

As mentioned earlier, migration at the time was common, a phenomenon that we operationalize for the machine learning algorithm through the distance between the county of the 1900 individual and that of the 1910 census potential match.

The algorithm is also provided with information on the 1900 individual's race, as well as whether this corresponds to the race of the 1910 potential match. The variable is intended to both capture underlying differences according to race in the ability to accurately link across censuses, as well as promoting matches when the 1900 individual and the 1910 potential match are recorded as belonging to the same race. Attempting to further distinguish between individuals with different underlying linkage probabilities, the algorithm is provided with information on the 1900 individual's region of birth, both domestic and foreign. Additionally, U.S. born individuals with (at least) one foreign born parent are indicated as being second generation immigrants. Lastly, and naturally only of relevance to the foreign born, we calculate the difference in the reported immigration year for the sample individual from the 1900 census and their potential match in 1910.

Training and implementing the algorithm

The approach selected for training the machine learning algorithm is similar to Feigenbaum (2016), however relying on a logistic rather than a probit regression model, as we found its performance consistently providing superior stability. Based on model parameters calibrated on training data, the machine assigns the predicted probability of a match for each 1900 and 1910 census potential match. The key decision therefore becomes which thresholds to select when declaring links in the data. This pertains to the *predicted probability cutoff* (α) value, roughly representing the required estimated similarity of two census records, as well as the *relative probability cutoff* (β) value, measuring in relative terms how much better than remaining potential matches the highest probability match is required to be. This is determined by employing a train-test-split procedure[5] and cross-validation over a range of realistic values of both thresholds in order to identify the optimal cutoffs. Selecting the optimal cutoffs is not straightforward, however, as the performance parameters of interest in record linkage, precision (share of identified matches

[5] This refers to a process where the training data is split into two, and where one half is used to calibrate model parameters and the other half is used to compare the resulting declared matches to the matches identified when generating the training data. By changing the α and β thresholds, different matches will be declared, however, always compared to the same underlying "gold standard" data.

that are correct) and recall (share of matches in the underlying training data that are identified) move in opposite directions as the thresholds are adjusted.

We select α and β thresholds based on Matthew's Correlation Coefficient (MCC), designed to be especially advantageous for use with unbalanced two-class data (Chicco 2017). The MCC, outlined in Eq. (1) below, compares the predictions of the algorithm to all possible outcomes (true/false positives/negatives) and provides a single metric (ranging from -1 to +1) to select which thresholds to use for overall optimal performance. For each unique combination of threshold values within a plausible range[6], we repeat the train-test-split procedure ten times in order to calibrate a stable MCC value. In the MCC formula displayed in equation (1) below, TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negatives:

$$(1)\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The model used to calibrate thresholds is presented in Table A1 in the Appendix, along with variable means in the underlying training data. Note that this is the model output using the full set of training data, and not the output from any of the separate train-test-split runs. Based on the average MCC obtained for each combination of α and β, we were able to identify the threshold values yielding the overall best performance, amounting to an MCC of 0.89[7], associated with a precision of 0.90 and a recall of 0.87.

Having obtained the thresholds, we proceed to implement the first stage of the linking procedure. For this exercise, we randomly selected 100,000 men from the 1900 full count census, linking them to potential matches following the same criteria as when generating the training data. Using the point estimates presented in Table A1 to predict the probability of each 1900-1910 potential match, subjected to the previously identified α and β thresholds, the algorithm yields 46,342 unique matches, translating to a linking rate of 46.3 percent.

Step II: Linking remaining household members

One of the major challenges associated with the feasibility of any attempt at record linkage is its considerable computational requirements. Therefore, applying blocking criteria to limit the universe of potential matches is typically implemented, possibly at the expense of the exclusion of de facto links through reporting and digitization errors in characteristics such as name and state or year of birth. The second step of our linking procedure takes advantage of the high-confidence links that were made in the

[6] The range of threshold values for α and β were 0.02-1 and 1-3, respectively.
[7] This is obtained using a value of α of 0.26 and a β of 1.

first step, while at the same time relaxing restrictive blocking criteria. The links already obtained provide valuable and specific information regarding where to look for individuals in the immediate circle of the person who was successfully linked in step I. More specifically, consider a situation like that in Figure 3 below, where Michael Corcoran and his son, John M Corcoran (both highlighted in grey) are among the individuals who we attempt to link in Step I. In this hypothetical yet hardly unusual example, we were unsuccessful in linking the son, likely because his first name in 1910 is listed as Mike rather than John M. Also likely are situations where the year of birth reported in the 1910 census fails to be within the +/-3-year interval or where there is a change in the individual's state of birth.

- Figure 3 here

The logic of the second step is to use the 1910 household identifier as the primary blocking criterion when creating the universe of potential matches for all remaining (unlinked) family members. Since this restricts the population of potential matches (and thus also the computational requirements) to such a great extent, we are able to relax all other blocking criteria, including place of birth, year of birth, sex, and name similarity score. In fact, the only remaining blocking criterion is year of birth (+/- 10 years). In the creation of training data consisting of 4,000 individuals, slightly differing characteristics were generated for use by the linking algorithm. The considerably lower degree of complexity of these characteristics reflects the much more straightforward task of linking once being able to block on the household level. This new level of simplicity is also reflected in the 1900 individual only being linked to 2.7 potential matches on average, as compared to 82 potential matches in Step I. The population of linkable individuals is reduced to 3,776 as a result of the blocking criteria used, and we were able to manually link 61.4 percent (2,320 individuals).

Linking variables

Similar to the earlier outlined procedure, the algorithm in step II is trained using characteristics such as Jaro-Winkler first and last name similarity scores, whether the race and place of birth of the 1910 potential match lines up with the 1900 census individual, and similarity in year of birth. Overall, however, a much more succinct set of linking variables are used, also reflecting the greater ease for the algorithm to confidently declare matches when the universe of potential matches is limited to the household.

While the incorrect enumeration or digitization of the individual's sex remains an unusual phenomenon in the data, it nevertheless does occur. In fact, in the 1900 census, there are over 100,000 people who have a gender that doesn't match their relationship to the household head, such as a daughter who is listed as male. As a consequence, a variable indicating whether the reported sex of the 1900 individual and the 1910 potential match is the same was created, naturally with the expectation that a mismatch on this characteristic will lower the predicted probability for any potential match. As indicated earlier, a more common

occurrence is that the recorded name(s) of an individual change over time, for example from John M in 1900 to Mike in 1910. In order to capture this possibility, categorical variables measure whether the first and middle name initials in 1900, respectively, matches the first or middle name initial of the 1910 potential match. As a consequence, for the John M to Mike potential match, the variable intends to signal the presence of evidence in favor of a link, whereas this would not be the case for the John M to Phil combination of records.

The two last linking characteristics we operationalize capture the individual's role within the household as well as the presence of potentially competing matches. Firstly, we believe that the individual's position within the household is characterized by varying degrees of volatility. More specifically, whereas a household head or spouse is likely to find themselves in a similar household a decade later, this is substantially less likely to be the case for a lodger or for a teenage child. Net of this, we believe that a 1900 individual and a 1910 potential match who occupy the same position within the household is a signal promoting a match. Lastly, we model the likelihood of incorrectly declaring a link by capturing the highest Jaro-Winkler score of a competing match in the household. The logic is perhaps most easily understood through an example like that displayed in Figure 4 below. In this situation, assume that we successfully linked the father, Stephen Frye, across the censuses and now turn to linking the remaining family members. In this case, the older of the two sons, Jonah, died in the intercensal period. Due to the high name similarity score between Jonah and Jonathan (JW=0.925), it is not, however, unlikely that many algorithms would have linked both Jonah and Jonathan in 1900 to the same 1910 record. As a consequence, both links would have been removed in the post-processing phase, as they represent duplicate matches. In an attempt to avoid this loss of data, for each 1900 individual, we obtain the highest name similarity score between the 1910 potential match in question and the remaining 1900 household members of that individual. In this case, for the Jonah 1900 record, while the name similarity score of Jonah and Jonathan (the 1910 potential match) amounts to 0.925, the corresponding value on the variable capturing the name similarity score of the most likely competing match amounts to 1. Consequently, the higher the latter score, the more likely it is that the potential match actually is another individual in the same household.

- Figure 4 here

Training and implementing the algorithm

The procedure used to calibrate the linking algorithm for Step II is the same as for Step I. Again, we opt for a logistic regression model when we calibrate the algorithm as well as which thresholds to use to optimize performance, performing ten loops over relevant threshold values to obtain stability. The estimated model

parameters, along with variable means, are presented in Table A2, Appendix, corresponding to a precision of 96.2% and a recall of 97.0% at a maximum MCC value of 0.96[8].

Implementing the second step is considerably less computationally demanding due to the ability to block on household identifier. The approximately 46,000 links from Step I yield 237,000 unlinked 1900 household members–both men and women–to be linked to the 1910 census in Step II. Despite the substantial number of individuals that we aim to link, the fact that the 1910 universe of potential matches is limited to household(s) to which their relatives belonging to the already confirmed matches have been linked limits the number of potential matches. Consequently, if a given linked individual is from a 1900 household where two members have been linked to *different* 1910 households in Step I, both these households will represent the universe of potential matches. Applying the parameter estimates and calibrated thresholds to the universe of potential matches, this results in another 104,932 matches, for a total of 151,274 linked men and women.

Performance comparison

The linking procedure introduced in this paper introduces two main innovations to automated record linkage: namely the algorithm's systematic use of a more extensive set of characteristics as well as implementing a two-step procedure where we are able to leverage the existence of confident links to substantially relax typically-used blocking criteria and thereby allow previously undiscoverable links to be found. It is, however, unclear how this approach performs compared to other methods of record linkage that are commonly encountered in the social scientific literature. We therefore proceed to implement both the Feigenbaum (2016) probabilistic record linkage method as well as the Abramitzky-Boustan-Erikson (2019) method (ABE), in order to see how they compare. We will limit the comparison to the links obtained in Step I, by applying aforementioned methods in linking the same 100,000 men, comparing the linked populations in terms of i) linkage rates, ii) precision, and iii) representivity.

When implementing the method proposed by Feigenbaum (2016), we use the same training data from Step I of the approach described in this paper[9]. In order to replicate the method in question as closely as possible to the original, however, we calibrate the performance thresholds using a probit estimator and with the same linking variables as Feigenbaum uses. Again, we select the α and β thresholds yielding the optimized performance based on MCC[10]. The ABE algorithm was implemented using the script provided by the

[8] Values used for α and β are 0.41 and 1.1, respectively.
[9] Parameter estimates are presented in Table A3, Appendix.
[10] The α and β values used were 0.08 and 1.7, resulting in an optimum MCC of 0.72. We suspect the training data used by Feigenbaum (2016) is characterized by a higher linkage rate than ours, despite the data that is linked being

authors[11], using first name, last name, and birthplace as the exact match variables. Year of birth information is also used, first looking for exact matches, subsequently expanding the search to individuals with up to a two-year difference. NYSIIS standard names were also used to compare among potential matches.

Linkage rate

Table 1 illustrates the number of individuals out of the sample population of 100,000 men that were confidently linked across the methods evaluated in this paper. Beginning with Step I of our method, 46,342 men, or 46.3 percent, are linked across the censuses. The Feigenbaum and ABE linking methods yield a substantially lower number of confirmed links, with the ABE method yielding 26,500 links and the Feigenbaum method yielding 28,400 links.

While the methods confidently link differing shares of the sample population, another relevant consideration pertains to the degree to which the methods–when declaring a match–come to the same conclusion. The agreement rate is thus calculated conditionally on both compared methods declaring a link for a given 1900 census individual. Across methods, the agreement rate is high, from 87 percent when comparing our method to Feigenbaum to 93 percent when comparing ABE with Feigenbaum.

- Table 1 here

Link accuracy

Despite the algorithms, to a large extent, being in agreement when declaring a match, a investigating how they perform when disagreeing will further our understanding of the advantages and disadvantages associated with selecting one approach over another. A frequently relied upon measurement, *precision*, is not independent of the process, as it is directly derived from the training data used to calibrate the linking algorithm. Instead, we make use of an extensive set of links from the Family Tree at familysearch.org. This comparison data is described in more detail in Price et al. (2019) and consists of pairs of records that have been attached to individual profiles on a genealogical website. This provides us with an excellent and unbiased way to investigate the *accuracy* of our declared links, as the assessment is independent of all methods that are being compared. It, however, needs to be underlined that while the database contains a very large number of high-confidence links that we can double check our declared links against, it does not cover the entire population. For example, out of the 1900 individuals who were successfully linked using Step I of our linking procedure, about 5% are covered by the database used to crosscheck performance. As

recorded 25 years apart. We believe our generally less impressive precision and recall values when calibrating our algorithm is linked to this difference in the training data.

[11] https://ranabr.people.stanford.edu/sites/g/files/sbiybj5391/f/abe_basic_approach_1.pdf

a result, precision estimates are based on the subset of individuals in the 1900 census who are successfully linked by each method that overlap with the Family Search database[12].

- Table 2 here

Table 2 illustrates that all methods similarly produce links that are of high quality in terms of agreement with the Family Search database. Among sample individuals from the 1900 census that are present and linked to a 1910 census record in the Family Search database, the accuracy across the three models ranges from a low of 87% (13% of declared matches do not agree with the Family Search database) for the Feigenbaum method to a high of 98% for the method described in this paper. Thus, despite linking a substantially higher proportion of the original sample of 100,000 randomly selected males from the 1900 full count census, the procedure outlined in this paper is able achieve a higher overall accuracy than both of the other algorithms.

Delving a bit deeper into the overall accuracy statistic unsurprisingly reveals a further increase when two algorithms are in agreement regarding the declared link. More specifically, when our algorithm declares the same link as either the Feigenbaum or ABE method, the Family Search database suggests an accuracy of 99 percent. Arguably of greater interest is the extent to which declared matches are correct across the methods when they arrive at different conclusions. Beginning with matches only declared by our algorithm, also representing the single most common category, the agreement with the Family Search database is only about one percentage point lower. In contrast, the decline in the agreement rate when examining matches only made by the Feigenbaum or the ABE algorithm is quite significant. Only in about 20 and 60 percent of the cases, respectively, does the match correspond with that in the Family Search database. Lastly, and further evidence to support our method's improved ability to accurately distinguish between potential matches, in the cases where both our and the Feigenbaum or ABE algorithm declares a match for a 1900 individual but disagrees on which 1910 individual represents the true match, at least 90 percent of such cases result in our link corresponding to the Family Search database.

Representivity

The evidence we have presented suggests that the method introduced in this paper performs better than the other evaluated methods, with respect to both accuracy and linkage rate. A nontrivial consideration is, however, to what extent a linked sample resulting from any method of automated record linkage is able to

[12] Manual precision checks of all categories of cases presented in Table 3 but not covered by the Family Search database are provided in the Appendix, Table A4. While the manual check consistently suggests lower accuracy than the numbers provided by the Family Search database, a large internal consistency between the performance of respective linking algorithm remains.

reflect the underlying 1910 populations, as strong selection mechanisms in the process of record linking may result in linked samples that are not representative of the population of interest. As far as the method outlined in this paper is concerned, this is may be a particularly relevant concern, since the household level features used to conduct linking may disproportionately link individuals experiencing household level and geographical stability over time. In order to investigate to what extent this appears to be the case, Table 3 presents each linked sample's composition according to a range of characteristics, comparing them to a random subsample extracted from the 1910 full-count census, evaluating differences in means through t-tests. The comparison population is restricted to men, as well as to individuals 7 year of age and older and–if foreign born–with a time of arrival no later than 1900, serving to restrict the comparison population to individuals who were credibly present in the 1900 sample population, given the blocking criteria imposed.

- Table 3 here

The linked populations resulting from all tested methods diverge from the comparison population across virtually all examined characteristics. Across all three linking methods, the resulting 1910 populations are older, more likely to be white, U.S. born and residing in a rural area, in addition to being characterized by a higher socioeconomic status than the comparison sample population. This is generally consistent with Bailey et al. (2019), who, in their evaluation of several of the most commonly used automated record linkage methods, note that "no method consistently produces representative samples". Of potentially greater importance are differences across methods in their (in)ability to generate linked populations reflecting certain 1910 population characteristics. As far as our method is concerned, Table 3 reveals that the Step 1 linked population contains an overrepresentation of individuals residing with parents as well as a share of lifetime migrants that is lower than in the 1910 population that it should reflect. Should one wish to analyze a population that more closely resembles the comparison 1910 population, the data thus suggest both the Feigenbaum and the ABE method offer a better ability to do so. There are, however, important implications linked to differences across methods' accuracy, since we know that links generated by the Feigenbaum/ABE methods are associated with considerably more noise, through incorrect links. The issue is further illustrated by Table 4, showing differences in means across the linking methods, depending on whether observations were labeled as accurate links by the Family Search database. Beginning with our method, there are few statistically significant differences in means between the two categories, suggesting that–for example–the proportion of incorrect links among the foreign born does not differ from the proportion among the native born.

- Table 4 here

Turning to the alternative methods, important differences between the characteristics of correctly and incorrectly linked individuals emerge. Specifically, for both the Feigenbaum and the ABE methods, incorrect links are disproportionately found among individuals who are black, of low socioeconomic status, living in an urban area and are lifetime migrants. Consequently, while these methods yield populations that in some respects more closely resemble the comparison 1910 population, our evidence suggests that this comes at the price of higher errors in the linkage of individuals possessing these very characteristics.

Step II Links

Thus far, only the first step of our linking procedure has been evaluated, as only it can directly be compared to the Feigenbaum and ABE linking procedures. We now proceed to evaluate the complete set of links generated by the approach introduced in this paper. As previously reported, Step II links include remaining household members of individuals that are successfully linked in Step I, benefiting from being able to very narrowly define the universe of potential matches. Additionally, while we restricted the population linked in Step I to men, in Step II we are able to link a large population of women, primarily children and spouses. Table 5 illustrates that the additional approximately 105,000 individuals that are linked in Step II are characterized by a similarly high agreement rate compared to the Family Search database as in Step I, amounting to 98 percent.

- Table 5 here

While Step I was developed to offer an alternative to already existing methods for linking historical records, the addition of Step II was primarily developed to be a tool useful for efforts to link complete-count populations or for research questions focusing on the household as a unit of analysis. The ability to isolate the universe of potential matches to such a great extent allows for the confident identification of links even when there is nontrivial discrepancy of traditionally-used blocking criteria, such as place or year of birth. Resulting from the exercise of this paper, Step II yields an additional population of 59,000 women and 46,000 men. Table 6 compares the male (Step I+II) and female (Step II) populations to a randomly selected population of 500,000 individuals from the 1910 census. Focusing initially on men, the added population from Step II makes the linked population slightly younger and from larger households than the 1910 comparison population. In addition, the proportion of the population that resides with parents increases further. For women, selection according to the investigated characteristics generally resembles those for men, with the linked population disproportionately being white, not being a lifetime migrant and residing with a family member and in a larger household than the average woman from the 1910 census.

- Table 6 here

Conclusions

A large empirical literature has emerged from numerous efforts to link individuals across historical records using various forms of deterministic and probabilistic machine learning techniques. This paper is borne out of an expectation that the use of an expanded set of characteristics throughout the linking process yields higher precision and linkage rates, despite potential shortcomings in terms of stronger selection mechanisms. More specifically, we propose the systematic use of information on household members and available contextual characteristics, both to maximize precision and to increase the ability to declare matches. We conclude that not only does our method yield a considerably higher number of links, but also with a higher degree of accuracy than the methods with which we compare. At the same time, it needs to be underlined that our method does yield a population that underrepresents lifetime migrants—a significant concern. It is also true, however, that there appear to be disproportionately more false links made by the ABE and Feigenbaum methods among the groups that are known to be more difficult to link, including lifetime migrants and non-whites. The consequences of sample selection, as well as differing rates of linking errors by subgroups associated with substantive research questions, are important and need to be explored in greater detail in future research.

The approach used in this paper is part of the process of eventually creating a longitudinal panel dataset that includes everyone that lived in the United States between 1850 and 1940: the Multigenerational Longitudinal Project. Consequently, the method is designed to be straightforwardly modifiable to link censuses besides 1900-1910, either by using the same training data or after generating data uniquely suited for the censuses in question. Our method achieves very high match rates and precision, but it is likely that future access to additional records (such as birth, marriage, death, and administrative records) will make it possible to achieve even higher match rates. Efforts like the Longitudinal, Intergenerational Family Electronic Micro-Database (Bailey 2018) and the Census Longitudinal Infrastructure Project (Massey et al. 2018) point in the direction of using multiple record types to link censuses. In addition, it is likely that combining machine learning with traditional genealogical approaches could identify linkages that are simply impossible to find using only one tool or the other (Price et al. 2020). Limiting linking approaches to immutable characteristics is likely to come at the cost of blocking or capping the match rates that are possible and increasing the number of false positives that are used to answer research questions.

Linkage strategies based on immutable characteristics that have dominated the literature for the past two decades highlighted the potential for selection bias in record linkage, and such linked datasets may still be preferred for some research questions. The approach in this paper is an effort to provide a complementary

approach that has its own set of advantages. Modern access to the full-count census records and fast computing makes it possible for researchers to test the conclusions of their analysis using both types of linked samples. In addition, the trade-off between the two approaches will only be present in the short-run, while the linkages of people across census records remains incomplete. It is likely that combining multiple methods will allow the state of the art to move more quickly to the eventual goal of linking the great majority of the population across all of the US census records, at which point concerns about selection bias will be greatly reduced.

The exercise presented in this paper is limited to only linking a small subsample of the underlying population from the 1900 full-count census to records from the 1910 census. The linking algorithm, available for download at ipums.org., was developed using the statistical software Stata (MP, version 14) and shares other methods' challenges when it comes to scaling up the populations that one wishes to link. Unsurprisingly, the multitude of linking features employed by our method implies that the running time as well as the storage space and memory required exceeds that of the Feigenbaum method due to the latter's relative simplicity. In order to overcome these performance issues, *hlink* has been developed by the Multigenerational Longitudinal Panel project. *hlink* uses a distributed processing engine to spread the workload across a cluster of computers, increasing the efficiency and the speed with which linking can be performed. As an example, linking the full 1900 population to the 1910 full-count census, yielding 30 million confirmed links, required a running time of 65 hours. The program, built on top of Apache Spark, will be released in the near future and is designed for the user to tailor the procedure outlined in this paper for other datasets, choosing which linking variables to include or exclude, as well as employing various linking methodologies, thereby further pushing the frontier of record linkage of individual level data.

References

Abramitzky, R., Boustan, L., and Eriksson, K. (2014): A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy* 122, 3, 467-506.

Abramitzky, R., Mill, R., and Pérez, S. (2018): Linking Individuals Across Historical Sources: A Fully Automated Approach. *National Bureau of Economic Research*, Working Paper no. 24324.

Abramitzky, R., Boustan, L.,Eriksson, K., Feigenbaum, J., & Perez, S., (2020): Automated Linking of Historical Data. *Journal of Economic Literature*, forthcoming.

Alexander, R., and Ward, Z. (2018): Age at Arrival and Assimilation During the Age of Mass Migration. *The Journal of Economic History* 78, 3, 904-937.

Bailey, Martha J. (2018): "Creating LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database." *University of Michigan Working Paper*.

Bailey, M., Cole, C., Henderson, M., and Massey, C. (2020): How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth. *Journal of Economic Literature,* forthcoming.

Beach, B., Ferrie, J., Saavedra, M., and Troesken, W. (2016): Typhoid Fever, Water Quality, and Human Capital Formation. *The Journal of Economic History* 76, 1, 41-75.

Blumin SM. (1976): *The Urban Threshold: Growth and Change in a Nineteenth-Century American Community*. Chicago: Univ. Chicago Press

Collins, W., and Wanamaker, M. (2015): The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. *The Journal of Economic History* 75, 4, 947-992.

Curti M. (1959): *The Making of an American Community: A Case Study of Democracy in a Frontier County*. Stanford, CA: Stanford Univ. Press

Feigenbaum, J. J. (2016): Automated Census Record Linking: A Machine Learning Approach.” W*orking Paper*.

Ferrie, J. P. (1996): A New Sample of Americans Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript. *Historical Methods* 29, 141-156.

Goeken, R., Huynh, L., Lenius, T.A., and Vick, R. (2011): New Methods of Census Record Linking. *Historical Methods* 44, 7-14.

Hopkins R. (1968): Occupational and Geographical Mobility in Atlanta, 1870–1890. *The Journal of Southern History*, 34, 2, 200–13.

Knights PR. (1971): *The Plain People of Boston, 1830–1860: A Study in City Growth*. New York: Oxford Univ. Press

Malin J.C., (1935): The Turnover of Farm Population in Kansas. *The Kansas Historical Quarterly*, 4, 164–87

Massey, Catherine G. "Playing with Matches: An Assessment of Accuracy in Linked Historical Data." *Historical Methods* 50, no. 3 (2017): 129-43.

Massey, Catherine G.; Katie Genadek; Trent Alexander; Todd Gardner; and Amy O'Hara. "Linking the 1940 U.S Census with Modern Data." Historical Methods, 51, no. 4 (2018): 246-257.

Price, J., Buckles, K., Van Leeuwen, J., and Riley, I. "Combining Family History and Machine Learning to Link Historical Records." NBER Working paper #26227, 2019.

Ruggles, S., Fitch, C., and Roberts, E. 2018. "Historical Census Record Linkage." Annual Review of Sociology 44: 19-37

Table 1: Linkage rate, by method

	Step I	Feigenbaum	ABE
Confirmed links	46,342	28,404	26,459
Linkage rate	46.3	28.4	26.5
Agreement rate among declared matches:			
Step I	-	86.7	90.7
Feigenbaum		-	92.6
ABE			-

Table 2: Accuracy, by method

	Step I	Feigenbaum	ABE
Confirmed links	46,342	28,404	26,459
Agreement rate with Family Tree database	*97.7*	*86.5*	*93.4*
Identical links declared by both algorithms		19,038	18,597
		98.5	*99.0*
Only linked in Step I		24,372	25,848
		97.0	*96.5*
Only linked by Feigenbaum/ABE		6,434	5,965
		18.9	*62.2*
Linked by both algorithms, different links		2,932	1,897
Step I		*95.0*	*90.7*
Feigenbaum/ABE		*0.4*	*5.4*

Table 3: Representivity, by method

	Census	Step I		Feigenbaum		ABE	
Age	31.57	33.06	***	35.29	***	32.71	***
7-20	0.33	0.33		0.28	***	0.32	***
21-45	0.44	0.41	***	0.43	**	0.44	
46-60	0.15	0.17	***	0.17	***	0.15	
61+	0.07	0.09	***	0.12	***	0.09	***
Race							
White	0.88	0.93	***	0.93	***	0.92	***
Black	0.11	0.07	***	0.07	***	0.08	***
Other	0.01	0.00	***	0.00	***	0.00	***
U.S. born	0.87	0.89	***	0.90	***	0.90	***
Household size	6.02	5.85	***	5.74	***	5.85	***
1	0.03	0.01	***	0.03	***	0.02	***
2-3	0.21	0.20	**	0.23	***	0.22	***
4-6	0.42	0.44	***	0.43	***	0.43	***
7-10	0.27	0.29	***	0.26	***	0.27	**
11+	0.07	0.05	***	0.06	***	0.06	***
Resides with spouse	0.43	0.47	***	0.50	***	0.46	***
Resides with parent(s)	0.39	0.45	***	0.37	***	0.41	***
Married	0.46	0.48	***	0.51	***	0.48	***
Socioeconomic index*	23.46	24.07	***	24.58	***	23.89	**
1-9	0.14	0.12	***	0.12	***	0.13	***
10-14	0.19	0.21	***	0.22	***	0.20	***
15-25	0.14	0.15	***	0.14		0.15	**
26+	0.17	0.18	***	0.19	***	0.18	***
Missing	0.36	0.34	***	0.33	***	0.34	***
Resides in rural area	0.57	0.59	***	0.59	***	0.58	***
Lifetime migrant (only for U.S. born)	0.26	0.21	***	0.27	*	0.25	***
Region of residence							
New England	0.07	0.08	***	0.09	***	0.08	***
Mid-Atlantic	0.19	0.19		0.15	***	0.18	***
East North Central	0.20	0.23	***	0.21	***	0.22	***
West North Central	0.13	0.15	***	0.15	***	0.15	***
South Atlantic	0.13	0.12	***	0.14		0.12	***
East South Central	0.09	0.09	*	0.09	***	0.08	***
West South Central	0.10	0.08	***	0.09	***	0.09	***
Mountain	0.03	0.02	***	0.03	*	0.03	
Pacific	0.05	0.04	***	0.05	*	0.05	*
Observations	100,000	46,342		28,404		26,459	

Notes: * Mean SEI score only calcucated for individuals with non-missing values

*** p<0.01, ** p<0.05, * p<0.1

Table 4: Accuracy across methods, by 1910 characteristic

	Step I			Feigenbaum			ABE		
	Status of link compared with FamilySearch database								
	Agree	Disagree		Agree	Disagree		Agree	Disagree	
Age	35.40	33.52		35.19	37.95	***	34.14	37.00	**
Race									
White	1.00	0.99		0.99	0.94	***	1.00	0.91	***
Black	0.00	0.01		0.01	0.06	***	0.00	0.09	***
Other	0.00	0.00		0.00	0.00		0.00	0.00	
U.S. born	0.94	0.95		0.95	0.94		0.95	0.91	**
Household size	6.19	5.93		6.14	5.87		6.16	6.47	
Resides with spouse	0.59	0.60		0.57	0.58		0.55	0.52	
Resides with parent(s)	0.42	0.37		0.42	0.27	***	0.45	0.28	***
Married	0.60	0.63		0.58	0.60		0.56	0.56	
Socioeconomic index									
1-9	0.09	0.12		0.09	0.14	***	0.09	0.18	***
10-14	0.30	0.30		0.29	0.27		0.27	0.25	
15-25	0.14	0.12		0.14	0.13		0.15	0.12	
26+	0.16	0.12		0.15	0.18		0.15	0.15	
Missing	0.31	0.35		0.33	0.28	**	0.33	0.29	
Resides in rural area	0.70	0.70		0.70	0.63	***	0.69	0.61	**
Lifetime migrant (only for U.S. born)	0.27	0.27		0.27	0.37	***	0.26	0.40	***
Region of residence									
New England	0.08	0.04		0.09	0.07		0.09	0.04	***
Mid-Atlantic	0.14	0.12		0.11	0.14	*	0.12	0.15	
East North Central	0.28	0.21	*	0.28	0.17	***	0.28	0.23	*
West North Central	0.18	0.17		0.19	0.16		0.19	0.16	
South Atlantic	0.10	0.15		0.10	0.17	***	0.09	0.14	**
East South Central	0.08	0.08		0.07	0.10	**	0.07	0.08	
West South Central	0.07	0.07		0.07	0.08		0.06	0.08	
Mountain	0.04	0.09	*	0.05	0.04		0.04	0.07	
Pacific	0.05	0.07		0.05	0.06		0.05	0.05	
Observations in Family Tree database	4,804	113		2,599	406		2,752	195	

Notes: *** p<0.01, ** p<0.05, * p<0.1

Table 5: Accuracy, Step I and Step II

	Step I	Step II	Total
Confirmed links	46,342	104,932	151,274
Agreement rate with FamilyTree database	*97.7*	*98.0*	*97.9*

Table 6: Representivity, Step I and Step II combined by sex.

	Men			Women		
	Census	**Step I+II**		**Census**	**Step II**	
Age in 1910	31.76	30.85	***	31.19	32.42	***
7-20	0.33	0.41	***	0.34	0.37	***
21-45	0.44	0.35	***	0.44	0.38	***
46-60	0.15	0.17	***	0.14	0.19	***
61+	0.08	0.07	***	0.08	0.06	***
Race						
White	0.89	0.94	***	0.88	0.94	***
Black	0.11	0.06	***	0.11	0.06	***
Other	0.01	0.00	***	0.00	0.00	***
U.S. born	0.87	0.89	***	0.89	0.87	***
Household size	6.00	6.55	***	5.72	6.71	***
1	0.04	0.01	***	0.02	0.00	***
2-3	0.21	0.12	***	0.23	0.10	***
4-6	0.42	0.41		0.43	0.41	***
7-10	0.27	0.38	***	0.26	0.41	***
11+	0.07	0.08	***	0.06	0.08	***
Resides with spouse	0.43	0.39	***	0.45	0.47	***
Resides with parent(s)	0.39	0.56	***	0.39	0.49	***
Married	0.46	0.40	***	0.47	0.48	***
Socioeconomic index*	23.41	23.27		25.03	29.23	***
1-9	0.14	0.11	***	0.04	0.02	***
10-14	0.19	0.19	***	0.02	0.01	***
15-25	0.14	0.17	***	0.07	0.07	***
26+	0.17	0.16	***	0.05	0.05	***
Missing	0.36	0.37		0.81	0.85	***
Resides in rural area	0.56	0.59	***	0.54	0.57	***
Lifetime migrant (only for U.S. born)	0.26	0.19	***	0.24	0.20	***
Region of residence						
New England	0.07	0.08	***	0.07	0.08	***
Mid-Atlantic	0.19	0.19	**	0.20	0.20	**
East North Central	0.20	0.23	***	0.20	0.24	***
West North Central	0.13	0.16	***	0.13	0.16	***
South Atlantic	0.13	0.12	***	0.14	0.11	***
East South Central	0.09	0.09	***	0.09	0.08	***
West South Central	0.10	0.08	***	0.09	0.07	***
Mountain	0.03	0.02	***	0.03	0.02	***
Pacific	0.05	0.04	***	0.04	0.03	***
Observations	255,334	92,335		244,666	58,939	

Notes: * Mean SEI score only calcucated for individuals with non-missing values

*** p<0.01, ** p<0.05, * p<0.1

Figure 1: Sample potential match data, only displaying (theoretically) immutable characteristics

1900				1910			
Last name	First name	Middle name	Birth year	Last name	First name	Middle name	Birth year
Thomas	Arthur	P	1892	Thomas	Arthur	F	1891
Thomas	Arthur	P	1892	Thomas	Arthur		1890
Thomas	Arthur	P	1892	Thomas	Arthur	H	1892
Thomas	Arthur	P	1892	Thomas	Arthur	J	1892
Thomas	Arthur	P	1892	Thomsen	Arthur	H	1894
Thomas	Arthur	P	1892	Thomson	Arthur		1893
Thomas	Arthur	P	1892	Thompson	Arthur		1891
Thomas	Arthur	P	1892	Thompson	Arthur		1892

Figure 2: Sample potential match data, extended with father's first name

1900				1910				Father's first name	
Last name	First name	Middle name	Birth year	Last name	First name	Middle name	Birth year	1900	1910
Thomas	Arthur	P	1892	Thomas	Arthur	F	1891	Benjamin H	Benjiman H
Thomas	Arthur	P	1892	Thomas	Arthur		1890	Benjamin H	
Thomas	Arthur	P	1892	Thomas	Arthur	H	1892	Benjamin H	Edward J
Thomas	Arthur	P	1892	Thomas	Arthur	J	1892	Benjamin H	Adrew D
Thomas	Arthur	P	1892	Thomsen	Arthur	H	1894	Benjamin H	Christian
Thomas	Arthur	P	1892	Thomson	Arthur		1893	Benjamin H	Thos
Thomas	Arthur	P	1892	Thompson	Arthur		1891	Benjamin H	Edward J
Thomas	Arthur	P	1892	Thompson	Arthur		1892	Benjamin H	Soren

Figure 3: Using confirmed individual links to define universe of potential matches for Step II

1900					1910			
Last name	First name	Birth year	Role		Last name	First name	Birth year	Role
Corcoran	Michael	1875	Head	→	Corcoran	Michael	1875	Head
Corcoran	Judy	1880	Spouse		Corcoran	Judith	1878	Spouse
Corcoran	John M	1898	Son		Corcoran	Mike	1899	Son
Corcoran	Philip	1895	Son		Corcoran	Phil	1897	Son
Corcoran	Lydia	1984	Daughter		Corcoran	Lydia	1895	Daughter

Figure 4: Operationalization of competing household match variable

1900			1910		
Last name	First name	Birth year	Last name	First name	Birth year
Frye	Stephen	1862	Frye	Stephen	1862
Frye	Jonah	1896	Frye	Jonathan	1898
Frye	Jonathan	1899			

Table A1: Logit estimates, Step I

	β	s.e.	P>|z|	lower 95%	upper 95%	Variable mean
1900 individual and 1910 potential match first name Jaro-Winkler score	8.63809	1.875	0.000	4.964	12.313	0.88
1900 individual and 1910 potential match standardized first name Jaro-Winkler score	3.80119	1.365	0.005	1.127	6.476	0.88
1900 individual and 1910 potential match last name Jaro-Winkler score	23.50728	1.315	0.000	20.930	26.084	0.78
1900 individual and 1910 potential match first name Soundex score	0.60182	0.268	0.025	0.077	1.126	0.47
1900 individual and 1910 potential match last name Soundex score	1.24501	0.220	0.000	0.814	1.676	0.10
Birth year difference between 1900 individual and 1910 potential match						
0	ref					0.15
1	-0.38058	0.152	0.012	-0.679	-0.083	0.29
2	-1.53711	0.190	0.000	-1.909	-1.166	0.29
3	-3.23388	0.291	0.000	-3.803	-2.664	0.28
Place of birth of 1900 individual:						
New England	ref					0.05
Mid-Atlantic	0.23970	0.341	0.483	-0.429	0.909	0.13
East North Central	1.33702	0.301	0.000	0.748	1.926	0.16
West North Central	1.95898	0.295	0.000	1.381	2.537	0.13
South Atlantic	0.46402	0.300	0.122	-0.124	1.052	0.13
East South Central	1.55907	0.323	0.000	0.926	2.192	0.10
West South Central	1.91913	0.345	0.000	1.243	2.595	0.07
Mountain	1.40185	0.307	0.000	0.800	2.003	0.02
Pacific	1.21746	0.395	0.002	0.443	1.992	0.02
North America/UK/Ireland/Nordic	0.81646	0.411	0.047	0.010	1.623	0.13
Rest of Europe	1.86893	0.421	0.000	1.044	2.694	0.05
Rest of World	1.37674	0.477	0.004	0.442	2.312	0.02
1900 individual is not second generation immigrant	ref					0.76
1900 individual is second generation immigrant	0.27565	0.171	0.106	-0.059	0.610	0.24
Difference between 1900 individual's and 1910 potential match's immigration year is less than or equal to 5 years (includes natives)	ref					0.87
Difference between 1900 individual's and 1910 potential match's immigration year is 6-10 years	-0.61866	0.499	0.215	-1.597	0.359	0.04
Difference between 1900 individual's and 1910 potential match's immigration year is 11 or more years	-1.40227	0.414	0.001	-2.213	-0.592	0.09
1900 individual race: White	ref					0.90
1900 individual race: Non-white	1.47032	1.238	0.235	-0.955	3.896	0.10
1910 potential match is same race						
No	ref					0.12
Yes	4.09470	1.051	0.000	2.034	6.155	0.88
Interaction: 1900 id and 1910 potential match are non-white	-2.11316	1.259	0.093	-4.580	0.354	
Distance between county of residence of 1900 individual and 1910 potential match	-0.00202	0.000	0.000	-0.002	-0.002	630.5
Distance between county of residence of 1900 individual and 1910 potential match, squared	0.00000	0.000	0.000	0.000	0.000	

Number of potential matches for 1900 id	-0.00566	0.001	0.000	-0.007	-0.004	499.6
Number of potential matches for 1900 id, squared	0.00000	0.000	0.000	0.000	0.000	
Number of same name individuals in 1900 (+birth place and +/-3 years year of birth)	-0.02968	0.007	0.000	-0.043	-0.016	23.1
Number of same name individuals in 1900, squared	0.00003	0.000	0.000	0.000	0.000	
Less than or equal to one 1910 potential match with the exact same first and last name as the 1900 individual	ref					0.60
At least one 1910 potential match with the exact same first and last name as the 1900 individual	-2.73271	0.186	0.000	-3.097	-2.368	0.40
Father not present in household of 1900 individual and 1910 potential match	ref					0.72
Father present, no mismatch on place of birth, age or relationship to id	-2.38217	0.474	0.000	-3.311	-1.454	0.04
Father present, mismatch on place of birth, age or relationship to id	-3.34452	0.449	0.000	-4.225	-2.465	0.24
Father of 1900 individual and 1910 potential match first name Jaro-Winkler score	5.74609	0.525	0.000	4.718	6.774	0.10
Mismatch on birth place of father of 1900 individual and 1910 potential match	ref					0.49
Match on birth place of father of 1900 individual and 1910 potential match	0.96987	0.175	0.000	0.627	1.313	0.51
Mother not present in household of 1900 individual and 1910 potential match	ref					0.68
Mother present, no mismatch on place of birth, age or relationship to id	-2.90439	0.480	0.000	-3.845	-1.964	0.05
Mother present, mismatch on place of birth, age or relationship to id	-4.34448	0.442	0.000	-5.210	-3.479	0.27
Mother of 1900 individual and 1910 potential match first name Jaro-Winkler score	5.79514	0.550	0.000	4.717	6.873	0.14
Mismatch on birth place of mother of 1900 individual and 1910 potential match	ref					0.47
Match on birth place of mother of 1900 individual and 1910 potential match	0.51139	0.181	0.005	0.157	0.865	0.53
Spouse not present in household of 1900 individual and 1910 potential match	ref					0.81
Spouse present, no mismatch on place of birth, age or marriage duration	-3.59735	0.485	0.000	-4.548	-2.647	0.04
Mother present, mismatch on place of birth, age or marriage duration	-5.60078	0.517	0.000	-6.615	-4.587	0.14
Spouse of 1900 individual and 1910 potential match first name Jaro-Winkler score	7.88033	0.610	0.000	6.685	9.075	0.09
Middle initial of 1900 individual and 1910 potential match is present and matches	ref					0.02
Middle initial of 1900 individual and 1910 potential match is present and mismatches	-4.02053	0.310	0.000	-4.628	-3.413	0.18
Middle initial of 1900 individual or 1910 potential match is missing	-2.85310	0.215	0.000	-3.275	-2.432	0.80
1900 individual and 1910 potential match shares less than 5% of neighbors	ref					0.99
1900 individual and 1910 potential match shares at least 5% of neighbors	2.94884	0.183	0.000	2.590	3.308	0.01
1900 individual and 1910 potential match street name Jaro-Winkler score less than 0.9 (or missing)	ref					1.00
1900 individual and 1910 potential match street name Jaro-Winkler score at least 0.9	2.60080	0.737	0.000	1.156	4.046	0.00
Relative not present in household of 1900 individual and 1910 potential match	ref					0.93
Relative present in household of 1900 individual and 1910 potential match	2.72318	0.303	0.000	2.129	3.317	0.07
Unrelated household member not present in household of 1900 individual and 1910 potential match	ref					0.98
Unrelated household member present in household of 1900 individual and 1910 potential match	-0.28504	1.392	0.838	-3.013	2.442	0.02
Intercept	-39.67307	2.105	0.000	-43.798	-35.548	
N			221,388			
Pseudo R2			0.8812			

Table 2: Logit estimates, Step II

	β	s.e.	P>|z|	lower 95%	upper 95%	Variable mean
1900 individual and 1910 potential match first name Jaro-Winkler score	2.026	0.835	0.015	0.389	3.663	0.52
1900 individual and 1910 potential match standardized first name Jaro-Winkler score	4.427	0.836	0.000	2.788	6.065	0.53
1900 individual and 1910 potential match last name Jaro-Winkler score	4.338	0.559	0.000	3.243	5.434	0.87
Birth year difference between 1900 individual and 1910 potential match						
0	ref					0.10
1	-0.720	0.234	0.002	-1.179	-0.262	0.18
2	-1.813	0.274	0.000	-2.350	-1.276	0.11
3	-3.813	0.360	0.000	-4.519	-3.107	0.10
4	-3.504	0.408	0.000	-4.304	-2.705	0.09
5	-3.885	0.409	0.000	-4.687	-3.084	0.08
6-10	-4.862	0.330	0.000	-5.510	-4.215	0.34
1900 individual is foreign born	0.875	0.821	0.286	-0.734	2.485	0.08
Place of birth of 1900 individual is the same as for 1910 potential match	1.418	0.366	0.000	0.701	2.136	0.79
Interaction: 1900 individual is foreign born & Place of birth of 1900 individual is the same as for 1910 potential match	-0.868	0.893	0.331	-2.618	0.882	
Race of 1900 individual is the same as for 1910 potential match	0.841	1.444	0.561	-1.990	3.671	0.99
Sex of 1900 individual is the same as for 1910 potential match	3.026	0.310	0.000	2.418	3.634	0.59
Change in marriage duration (if married in both censuses) between 1900 individual and 1910 potential match is between 6-14 years	1.283	0.377	0.001	0.545	2.021	0.15
1900 individual is household head or spouse	ref					0.18
1900 individual is child or child-in-law	0.892	0.560	0.111	-0.206	1.989	0.72
1900 individual has other relationship to household head	1.054	0.625	0.092	-0.171	2.279	0.09
Relationship to household head of 1900 individual is the same as for 1910 potential match	2.361	0.485	0.000	1.410	3.313	0.73
Interaction: Relationship to household head of 1900 individual is the same as for 1910 potential match & 1900 individual is child or child-in-law	-1.314	0.619	0.034	-2.527	-0.100	
Interaction: Relationship to household head of 1900 individual is the same as for 1910 potential match & 1900 individual has other relationship to household head	-1.682	1.002	0.093	-3.646	0.282	
Highest Jaro-Winkler score of competing match in household, conditional on +/- 10 year birth year difference	-0.763	0.403	0.058	-1.553	0.026	0.45
Highest Jaro-Winkler score of competing match in household, unconditional	-2.953	0.474	0.000	-3.883	-2.023	0.69
No first name	ref					0.02
First name initial of 1900 individual matches first or middle name initial of 1910 potential match	0.698	1.375	0.612	-1.998	3.393	0.29
First name initial of 1900 individual mismatches both first and (if available) middle name initial of 1910 potential match	-1.576	1.353	0.244	-4.228	1.075	0.71
No middle initial	ref					0.64
Middle name initial of 1900 individual matches first or middle name initial of 1910 potential match	1.115	0.249	0.000	0.628	1.603	0.09
Middle name initial of 1900 individual mismatches both first and (if available) middle name initial of 1910 potential match	-0.545	0.218	0.012	-0.973	-0.118	
Intercept	-11.625	2.136	0.000	-15.811	-7.438	
N			10,101			
Pseudo R2			0.8965			

Table A3: Probit estimates, Feigenbaum machine learning model

	β	s.e.
First and Last name match	0.452	(0.0696)
1900 individual and 1910 potential match first name Jaro-Winkler score	5.774	(0.445)
1900 individual and 1910 potential match last name Jaro-Winkler score	7.402	(0.360)
Birth year difference between 1900 individual and 1910 potential match		
0	ref	
1	-0.234	(0.0429)
2	-0.894	(0.0565)
3	-1.356	(0.0742)
1900 individual and 1910 potential match first name Soundex match	0.034	(0.0804)
1900 individual and 1910 potential match last name Soundex match	0.532	(0.0643)
Number of potential matches for 1900 id	-0.004	(0.000189)
Number of potential matches for 1900 id, squared	0.000	(5.63e-08)
Multiple matches for first and last name	-1.307	(0.0688)
First letter of first name matches	0.695	(0.133)
First letter of last name matches	0.358	(0.0849)
Last letter of first name matches	-0.051	(0.0710)
Last letter of last name matches	0.316	(0.0581)
Middle initial match, if there is a middle initial	1.001	(0.0625)
Intercept	-14.590	(0.512)

Table A4: Results from manual accuracy checks of links, by category

	Step I	Feigenbaum	ABE
All links	*90*	*68*	*82*
Identical links declared by both algorithms		*96*	*96*
Only linked in Step I		*94*	*90*
Only linked by Feigenbaum/ABE		*14*	*32*
Linked by both algorithms, different links			
Step I		*92*	*78*
Feigenbaum/ABE		*0*	*8*

Note: For each category, 50 cases were randomly ectracted and manually evaluated using Ancestry.com. The handlinker had no knowledge to which category any given case belonged.