Advancing the Frontier for Linking Subgroups in US Census Records

Adrian Haws
Cornell University

Kasey Buckles
Notre Dame University

Joseph Price
Brigham Young University

Abstract

Recent technologies have rapidly enhanced the match rates and accuracy of historical US census links, cultivating rich opportunities for longitudinal and intergenerational research. However, accurately linking certain racial and ethnic minority groups remains significantly more challenging.. While machine learning-based methods perform better than rules-based methods, both types of linking methods have large disparities in match rates by race and ethnicity. We identify challenges in linking Black Americans, Native Americans, and immigrants from Italy, Yugoslavia, Mexico, and Japan, and describe four strategies that can be used to improve match rates. Our strategies are informed by a massive set of ground truth census links available through the Census Tree project.

The availability of full-count US census microdata between 1850 and 1950 and recent advances in record linking methods have made it possible to create massive datasets following individuals and families over time. These newly linked datasets (Abramitzky et al. 2021; Price et al. 2021; Helgertz et al. 2022; Bailey et al. 2023) have enabled researchers to explore new questions and to revisit critical debates using much larger sample sizes. However, the currently available census linkages are not always suitable for studying subsets of the population, which are the focus of many research questions. For example, we find that several racial and ethnic minorities are linked at less than half the rate of White US-born men. Studies of these groups may therefore be limited by imprecise statistical inference. Very low match rates (under 5 percent for some groups) also raise concerns about the degree of selection into the linked sample, which could bias estimates.

In this paper, we use millions of hand-linked census records, created by personal genealogists, to inform specific strategies for improving match rates for population subgroups within a machine learning framework. We focus on six subgroups: Black Americans, Native Americans, and immigrants from Italy, Mexico, Yugoslavia, and Japan. We plan to provide the code and data required for scholars to refine and expand our approach to subsets of the population that are relevant for their specific research.

We also demonstrate how to create additional matches for subsets of the population by combining automated linking methods with digital genealogical tools. This approach builds on strategies that have been used to link applicants of the Mother's Pension Program (Aizer et al. 2024a), participants in the Civilian Conservation Corps (Aizer et al. 2024b), and students who attended Harvard in the early 20th century (Michelman et al. 2022). Our methods can greatly

lower the cost of creating samples with high match rates such as the Early Indicators project (Costa et al. 2017).

Our work complements recent efforts to improve census linking match rates for under-represented groups, most notably for women who change their surnames when they marry (Althoff et al. 2023; Bailey et al. 2023; Buckles et al. 2024; Helgertz et al. 2022). Racial and ethnic minority groups have received comparatively less attention in the record linking literature, though a notable exception is Postel (2023) who develops methods for linking Chinese Americans and other individuals with character-based surnames. The tools we describe here can be used to improve linking rates for a wide range of minority populations that are of interest to researchers.

Data

The Family Tree

We use millions of hand-linked census records from the Family Tree at familysearch.org as the foundation of our automated census linking strategies. FamilySearch is a free genealogy platform that includes over 13 billion searchable historical records. Users of the FamilySearch platform collaborate on a shared, population-wide Family Tree which includes 1.64 billion profiles for deceased individuals. Each profile has biographical details of the deceased individual, as well as attached records such as birth, marriage, death, and census records. Because users can edit and build on each profile and its attached records, the Family Tree continually becomes more comprehensive and accurate.

A key innovation first introduced by Price et al. (2021) is that the census records attached to each profile on the Family Tree can be used to create a massive set of "ground truth" census

links to train a machine learning algorithm. When a deceased person's profile on the Family Tree is attached to records in the 1910 and 1920 US censuses, for instance, this pair of census records is an example of a true match. Some of these links would be very difficult even for trained research assistants to establish, because FamilySearch users often have private biographical information about the deceased person, such as maiden names and aliases. While users on FamilySearch may sometimes attach census records to the incorrect profile, the FamilySearch software helps users to find and fix these mistakes. In an audit of Family Tree links, where research assistants checked links from several linking methods without knowledge of the source, Buckles et al. (2024) report that 98 percent of Family Tree census links were found to be correct. Other researchers have used the Family Tree links as "ground truth" to measure the quality of their automated census links (Bailey et al. 2020).

The Family Tree provides a much larger training set than would be feasible by hiring research assistants or professional genealogists. Although a large training set is advantageous because it includes more variation in the types of links and allows a high-dimensional set of comparisons, the key advantage for our application is that even very small subgroups of the population are represented in the data. Table 1 shows the number of Family Tree links that are available for the full population and for the subgroups that we study. Our focus is on linking 1910 and 1920 census records, with nearly 31 million Family Tree links available. Each of the featured population subgroups has a sizable number of links, with 0.7 million links for Black Americans alone. Although just over 82,000 Japanese immigrants are recorded in the 1920 US Census, 654 "ground truth" links are still available.

The broad representation of the Family Tree is a critical strength for improving census linking for population subgroups. In addition to providing training data for machine learning

algorithms, the Family Tree dataset helps diagnose weaknesses in current linking approaches and develop solutions. We can learn why some links in the Family Tree are not found by automated methods, and why some automated links disagree with the Family Tree links. These data also inform how we prepare the census fields for linking (especially name strings) and which information about candidate links should be given to machine learning algorithms.

The Family Tree data is already publicly available and can be downloaded at censustree.org. The Census Tree dataset, described by Price et al. (2021) and Buckles et al. (2024), includes links from the Family Tree as well as links from automated methods (including from a machine learning algorithm trained on the Family Tree data). Most researchers use the Census Tree directly, but the Family Tree data can be obtained by selecting links where the "family_tree" variable is equal to 1.

Current Census Linking Methods

To evaluate shortcomings in linking population subgroups, we focus on four prominent historical census linking methods which include machine learning and rules-based algorithms. First, we examine the Census Tree – Machine Learning (Census Tree – ML) links, which are a subset of the Census Tree dataset (Price et al. 2021; Buckles et al. 2024). The Census Tree – ML links are created by using 114 million total links from the Family Tree to train a unique machine learning model for each combination of full-count census records between 1850 and 1940.

We also examine the full Census Tree, which includes links from the Family Tree, Census Tree – ML, additional machine learning links provided by the FamilySearch software, and automated links from the IPUMS Multigenerational Panel (Helgertz et al. 2022) and the Census Linking Project (Abramitzky et al. 2014; Abramitzky et al. 2021). The Census Tree includes a

process for adjudicating disagreements across sources when they occur. Both the full Census Tree and the Census Tree – ML links include men and women.

The IPUMS Multigenerational Longitudinal Panel (IPUMS MLP; Helgertz et al. 2022) uses a two-step machine learning approach, first linking individual men before linking individuals in their household regardless of sex. Links are created for 10-, 20-, and 30-year differences across census years from 1850 to 1950. One machine learning model is used to create Step I links across 10-year differences, using a set of 1,354 hand links between 1900 and 1910 as training data. The training set contains a total of 119 foreign-born men. Separate models for 20-year and 30-year differences are trained on hand links for 1880-1900 and 1900-1930.

The Census Linking Project was the first set of population-level US census links that were made available to researchers. This rules-based method, described by Abramitzky et al. (2012) and Abramitzky et al. (2021), generates census links for individuals who are unique across a small set of (presumably) immutable characteristics. Characteristics include names, sex, birthplace, birth year, and, in some cases, race. Because women's surnames often change when they marry, only men are included in these links. Publicly available links include six variations of this algorithm, intended to offer a trade-off between match rates and accuracy: "Exact Conservative", "NYSIIS Conservative", "NYSIIS Standard – Race", "NYSIIS Conservative – Race", "NYSIIS – Standard", and "Exact Standard". We evaluate links based on the "Exact Conservative" method because its estimated accuracy is most comparable with the IPUMS MLP and the Family Tree.

Match Rates for Racial and Ethnic Minorities

Figure 1 displays the fraction of men linked to the 1910 census using four census linking methods: the full Census Tree, Census Tree – ML, IPUMS MLP, and the Census Linking Project links. Match rates are shown for White, Black, and Native American individuals born in the US, as well as for the largest 23 immigrant groups in the 1920 census. The sample is restricted to men who are 10 years or older in 1920 and who immigrated before 1910 (if applicable). Nearly 80 percent of White men born in the US are linked in the full Census Tree, while 18 percent of immigrants from Japan are linked. Extreme differences in match rates are also evident across the other three methods. Census Tree – ML links 50 percent of English immigrants and 8 percent of Japanese immigrants; IPUMS MLP links 60 percent of US-born White individuals and 0.5 percent of Japanese immigrants; and the Census Linking Project links 34 percent of US-born White individuals and 0.04 percent of Lithuanian immigrants.

Group-level differences across match rates persist for large subsets of the population as well. Match rates for African Americans are between 53 percent (in the full Census Tree) and 15 percent (using the Census Linking Project). Linking methods with lower match rates are no more representative—the ratio of the African American match rate to the US-born White match rate in the full Census Tree is 0.68, and the corresponding ratio in the Census Linking Project is 0.43.

A Machine Learning Approach for Linking Key Subgroups

Low match rates for subgroups of the population, across census linking methods, presents a crucial challenge for empirical research. We leverage "ground truth" links from the Family Tree to explore why some groups are more difficult to link, and to suggest why some linking methods perform better on different dimensions. Using these insights, we outline a set of

strategies to improve match rates within a machine learning framework. We focus on six groups which may each have different reasons for lower match rates—African Americans, Native Americans, and immigrants from Italy, Mexico, Yugoslavia, and Japan.

The match rates shown in Figure 1 make it clear that both machine learning and rules-based methods underperform for many racial and ethnic minority groups. We develop strategies for improving machine learning algorithms, however, because machine learning generates overall higher match rates and because machine learning can fully leverage the training data available from the Family Tree. Some of the strategies we describe can also be applied to rules-based methods. While our remaining analysis focuses on linking men, they could be adapted to link women as well.

Our strategies correspond to four key steps of a machine learning process. See Price et al. (2021) and Helgertz et al. (2022) for more discussion of machine learning applications to historical census linking. These strategies are not all new, but we provide new evidence of how they can be combined and applied to advance the frontier of linking population subgroups.

Strategy 1 – Language-Specific Name Processing

Every automated linking algorithm requires a preliminary step of pre-processing name strings, because a person may not be recorded with the same name across census records. Differences can arise from the use of nicknames, middle names, or errors in enumeration, transcription, and digitization. Phonetic standardization is the most common method for pre-processing names.

Table 2 provides evidence that phonetic standardization can cause more immigrants in the Family Tree (linked between 1910 and 1920) to be considered an implausible match. Although 82 percent of Black Americans and 83 percent of Native Americans in the Family Tree match on

the NYSIIS phonetic encoding of first names, after converting common nicknames (e.g. Mike to Michael), only 79 percent of Italian immigrants and 75 percent of Mexican immigrants match. NYSIIS encoding is used to limit plausible links for both the Census Tree – ML links and the Census Linking Project. An extreme case is Japanese immigrants, who only match on standardized names 50 percent of the time. This means that only half of the Family Tree links for Japanese immigrants would be possible to link, before considering other limitations. Last names are less of a concern for immigrants from Italy, Mexico, and Yugoslavia, but only 44 percent of immigrants from Japan match on the NYSIIS encoding of last names.

Name pre-processing is not only a concern for limiting plausible links (also known as blocking), but it also matters for the quality of information input for machine learning algorithms. IPUMS MLP uses a much more flexible approach to compare first names, which considers 86 percent of Japanese immigrants as a plausible match. Two names are a plausible match if they share the same first letter or if any two-letter pairs match. However, it is difficult to determine which names are similar enough to be considered the same.

For Japanese immigrants, character-based name processing is a promising strategy. In her work on linking Chinese Americans, Postel (2023) demonstrates that match rates are three times higher and accuracy is improved when following a three-step approach to segmenting, ordering, and standardizing character-based names. Although this approach is most likely to help with Chinese Americans and other individuals with character-based names, it could also be helpful for populations with multiple surnames. For example, Mexican Americans often inherit surnames from both their father and mother, but census enumerators may inconsistently record these. A new approach could account for the possibility of the first surname being recorded as a given name, or for both surnames being present in one census and only one in another census.

Strategy 2: Expanding the Set of Candidate Links

Because it is infeasible to generate machine learning predictions for every possible pair of census records—there are 2.6 quadrillion combinations of records for men between the 1910 and 1920 US Census—machine learning methods will follow a blocking procedure to select plausible candidate links. The Census Tree dataset (Price et al. 2021; Buckles et al. 2024) blocking strategy requires that candidate links match on sex, birthplace, and the NYSIIS encoding of first and last names, and that the absolute difference in birth years is no greater than 3 years. IPUMS MLP – Step I (Helgertz et al. 2022) follows a similar strategy, except that plausible links includes those that match on either the first letter or any two adjacent letters (bigram) in the first and last name. (Step II is much less restrictive because only family members of Step I matches are considered). IPUMS MLP also requires that the first and last name have a Jaro-Winkler similarity score of at least 0.7.

There is an inherent trade-off in blocking—any restriction of candidate links will likely remove some true links. We provide evidence that common blocking restrictions are more limiting for some population subgroups. Price et al. (2021) reports that 66 percent of Family Tree links between 1910 and 1920 pass the blocking criteria for the Census Tree – ML links. However, in Table 2 we show that less than half of the Family Tree links for Black Americans, Native Americans, and immigrants from Mexico and Japan pass these blocking criteria. Limitations also vary by subgroup. For instance, Price et al. (2021) reports that 97 percent of Family Tree links have an absolute birth year difference of less than 3 years, but we show that only 77 percent of links for Mexican immigrants pass this condition. Other groups range from 85 to 90 percent.

The blocking criteria for IPUMS MLP – Step I suggests that blocking can be more efficient while remaining feasible. Up to 77 percent of Family Tree links are retained by blocking (for Italian immigrants). However, only 1 percent of links for immigrants from Yugoslavia are retained by blocking. The key difference is that the Census Tree – ML method permits links across any country definition in Europe, and Yugoslavia was created in 1918 after World War I.

We suggest three feasible ways to expand the set of candidate links for population subgroups. First, combining existing blocking methods is likely to retain more true links without including too many candidate links. In this scenario, the blocking rules would allow bigram matches within names as well as links across countries in central Europe. Second, blocking rules can be adjusted for specific subgroups, such as allowing larger birth year differences for immigrants from Mexico. Third, some subgroups are small enough to relax many blocking rules. The 1920 US Census records 47,546 men born in Yugoslavia who were 10 years or older and had immigrated before 1910. If each of these men were considered as a possible link to each of the 47.4 million men in 1910, there would be 2.3 trillion candidate links. However, limiting the set to the 2.4 million men born in central Europe lowers this to 0.1 trillion candidate links. Even if we followed a comparatively relaxed set of restrictions for birth year and name differences, the set of candidate links would be computationally feasible.

Strategy 3: Subgroup-Specific Feature Comparisons

Machine learning methods for census linking are currently designed to link everyone in a population. Feature comparisons (analogous to independent variables in a regression model) are chosen to allow the algorithm to learn what predicts a good match. For example, a feature for "absolute birth year difference" might allow the model to learn that an exact birth year match is

most often a true link in the ground truth data. However, these population-based features may not be ideal for specific subgroups.

The massive set of training data available from the Family Tree allows for a data-driven approach to selecting feature comparisons when linking subgroups. Table 2 shows that 15 percent of African American men in the Family Tree have a birth year difference greater than 3 years, but it is unclear whether this means that birth year differences should be less informative in a machine learning algorithm. We show in Figure 2 that birth year differences are more informative for younger African Americans, who are more likely to be literate. Among the nearly 358,000 African American men in our Family Tree sample, we show that men born in 1875 or earlier have a birth year difference of greater than 3 years on average. For individuals born between 1850 and 1910, the average birth year difference declines from about five years to about one year. This suggests that, for African American men in 1920, a useful feature to include in the machine learning model would be the interaction of age and the birth year difference. We also see that, for men born before 1900, reporting a birth year ending in 0 or 5 is associated with much more uncertainty in the birth year difference. Thus, another useful feature would be the interaction of an indicator for a birth year ending in 0 or 5 and the birth year difference.

Another example of subgroup-specific features is for immigrants' birthplaces. We can easily resolve the problem of Yugoslavia not matching to other birthplaces by defining a slightly more expansive blocking criteria—Yugoslavia can match to any country in central Europe. However, there are many candidate links that satisfy this criterion. In this case, a useful feature would be whether a birthplace of Yugoslavia in the 1920 census matches a birthplace of any of the territories that previous formed Yugoslavia in the 1910 census: Croatia, Slovenia, and Bosnia territories in the Austro-Hungarian Empire, and the Serbian Kingdom.

Strategy 4: Representative Training Data

The final strategy is to ensure that the machine learning model is trained on ground truth links from the key population subgroup, because this allows the model to learn the sets of characteristics (feature values) that constitute a likely match. Including subgroup-representative training data is only possible when the ground truth links are considered as candidate links (Strategy 2). Supervised machine learning models learn from both "true" and "false" links, and the "false" links are created by identifying candidate links (which pass the blocking criteria) that differ from the "true" link.

The size of the Family Tree makes it possible to find enough training data for many population subgroups that researchers may be interested in. Methods that create training links by hand are much less likely to represent key subgroups—IPUMS MLP uses 119 ground truth links for foreign-born men, which are unlikely to contain many links for any particular immigrant group. If the Family Tree does not contain enough training data for a specific application, the method we describe in the next section offers a way to generate additional ground truth links.

Combining Automated and Manual Linking

We have shown that match rates for population subgroups are often quite low, which introduces selection bias and limits statistical inference. Focusing on small groups could present an opportunity to overcome both of these challenges. While it is difficult to imagine hand-linking the full US population between 1910 and 1920, it may be much more feasible to link nearly all individuals in a small and well-defined subgroup. Previous efforts such as the Early Indicators project (Costa et al. 2017) and the Guild of One-Name Studies (Clark 2023) have created detailed linkages for Union Army veterans and for rare surnames in England. Here we introduce

a method that could make comprehensive record linkages quicker and more affordable by using traditional genealogical tools to hand-link records that were not linked by automated methods.

The Early Indicators project created record linkages by generating a family tree for each Union Army veteran on Ancestry.com (Costa et al. 2017). Like many genealogical websites, Ancestry provides powerful search tools, hints to possible matching records, and a massive collection of digitized historical records. Digitized records often provide information about individuals that enable linking them to multiple census records. For example, a birth record can show a person's first and middle name, which makes it possible to link them to their first name in one census and their middle name which they use as a preferred name in another census. Thus, while the linking task is to connect records A and B, the best approach may be to first link A to C before linking C to B. Information from multiple sources can also help to avoid false matches.

FamilySearch provides a similar set of search tools, record hints, and a collection of over 13 billion digitized historical records. The collaborative Family Tree, which is free to use, contains a large set of existing record linkages. Several papers demonstrate the potential for using FamilySearch to increase match rates. Aizer et al. (2024a) created a linked sample of 16,000 women who applied for the Mothers' Pension Program. Aizer et al. (2024b) created a linked sample of nearly 24,000 young adults who participated in the Civilian Conservation Corps. Michelman et al. (2022) created a linked sample of 14,000 students who attended Harvard in the early 20th century.

Using FamilySearch to increase the match rate for a smaller sample involves the following steps. First, a researcher uses the FamilySearch API to identify individuals in their starting dataset who already have a profile on the Family Tree. Second, the researcher creates a new profile for each person in the sample who is not yet represented on the Family Tree. This

can either be done by hand, using the "Add Unconnected Person" function on the FamilySearch website, or it can be automated (with permission from FamilySearch) by using an application created by the Record Linking Lab at Brigham Young University. Third, for the profiles that were already on the Family Tree, a researcher can check which of those profiles are linked to a record in the target dataset, thus providing a set of links that are already complete. Fourth, for the profiles that were added to the Family Tree or do not already have a link to the target dataset, the researcher can use record hints and search tools on FamilySearch to create additional links.

Links that were previously available or that the researcher added to the Family Tree can then be combined with automated links to attain a higher match rate than would otherwise be possible. The fourth step can also be expanded by finding matches using the search tools or publicly available family trees on other genealogical websites such as Ancestry, MyHeritage, or FindMyPast. Any record matches found on other websites can then be found and linked to the FamilySearch profile. The process may involve linking to other record collections to obtain more accurate information on birth dates and places, full names, maiden and married names, death dates and places, and places of residence. This information can help identify the correct match in the target dataset, identify deceased people who are impossible to link, and select the best of multiple possible matches.

The hand-linking process is similar to what was used to create the intergenerational linkages for the Early Indicators Project as described in Costa et al. (2017). The authors note that the search tools on Ancestry.com allowed them to achieve high linkage rates across census records for Civil War veterans and their descendants. Linkages from the Early Indicators Project are often described as the "gold standard" of linked historical data. The advantage of creating links on the collaborative Family Tree at FamilySearch, rather than on isolated family trees

created by Ancestry.com users, is that many links have already been created on the Family Tree by individuals working on their own family history.

Conclusion

Recent advances in automated census linking methods have made it possible to answer many research questions in economics, sociology, political science, and other fields. However, research conclusions drawn from these linked samples can be biased if subgroups of the population are systematically underrepresented. For example, Ward (2023) demonstrates that intergenerational mobility estimates can differ greatly when samples include Black Americans. The ability to study small groups of the population, such as specific immigrant groups, is also limited by lower match rates.

This paper highlights the role of genealogical data in improving match rates for population subgroups. We use a subset of 114 million links created by users of a collaborative Family Tree to expand census links for Black Americans and several immigrant groups. The Family Tree includes training data for small subgroups which can be used to inform a flexible machine learning process. We also demonstrate how automated record linking can be combined with manual methods to attain very high match rates for subgroups in the United States.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration." *Journal of Political Economy* 122, no. 3 (2014): 467-506.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. "Automated linking of historical data." *Journal of Economic Literature* 59, no. 3 (2021): 865-918.

Aizer, Anna, Sungwoo Cho, Shari Eli, and Adriana Lleras-Muney. "The Impact of Cash Transfers to Poor Mothers on Family Structure and Maternal Well-Being." *American Economic Journal: Applied Economics* 16, no. 2 (2024): 492-529.

Aizer, Anna, Nancy Early, Shari Eli, Guido Imbens, Keyoung Lee, Adriana Lleras-Muney, and Alexander Strand. "The Lifetime Impacts of the New Deal's Youth Employment Program." *The Quarterly Journal of Economics* (2024): 2579-2635.

Althoff, Lukas, Harriet Brookes Gray, and Hugo Reichardt. "The Missing Link: Women and Intergenerational Mobility." 2023. *Stanford University Working Paper* (2024).

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. "How well do automated linking methods perform? Lessons from US historical data." *Journal of Economic Literature* 58, no. 4 (2020): 997-1044.

Bailey, Martha, Peter Z. Lin, A.R. Shariq Mohammed, Paul Mohnen, Jared Murray, Mengying Zhang, and Alexa Prettyman. "The creation of LIFE-M: The longitudinal, intergenerational family electronic micro-database project." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56, no. 3 (2023): 138-159.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley E.B. Wilbert. "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project." No. w31671. *National Bureau of Economic Research,* 2024.

Clark, Gregory. "The inheritance of social status: England, 1600 to 2022." *Proceedings of the National Academy of Sciences* 120, no. 27 (2023): e2300926120.

Costa, Dora L., Heather DeSomer, Eric Hanss, Christopher Roudiez, Sven E. Wilson, and Noelle Yetter. "Union Army veterans, all grown up." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50, no. 2 (2017): 79-95.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J. Thompson, Steven Ruggles, and Catherine A. Fitch. "A new strategy for linking US historical censuses: A case study for the IPUMS multigenerational longitudinal panel." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55, no. 1 (2022): 12-29.

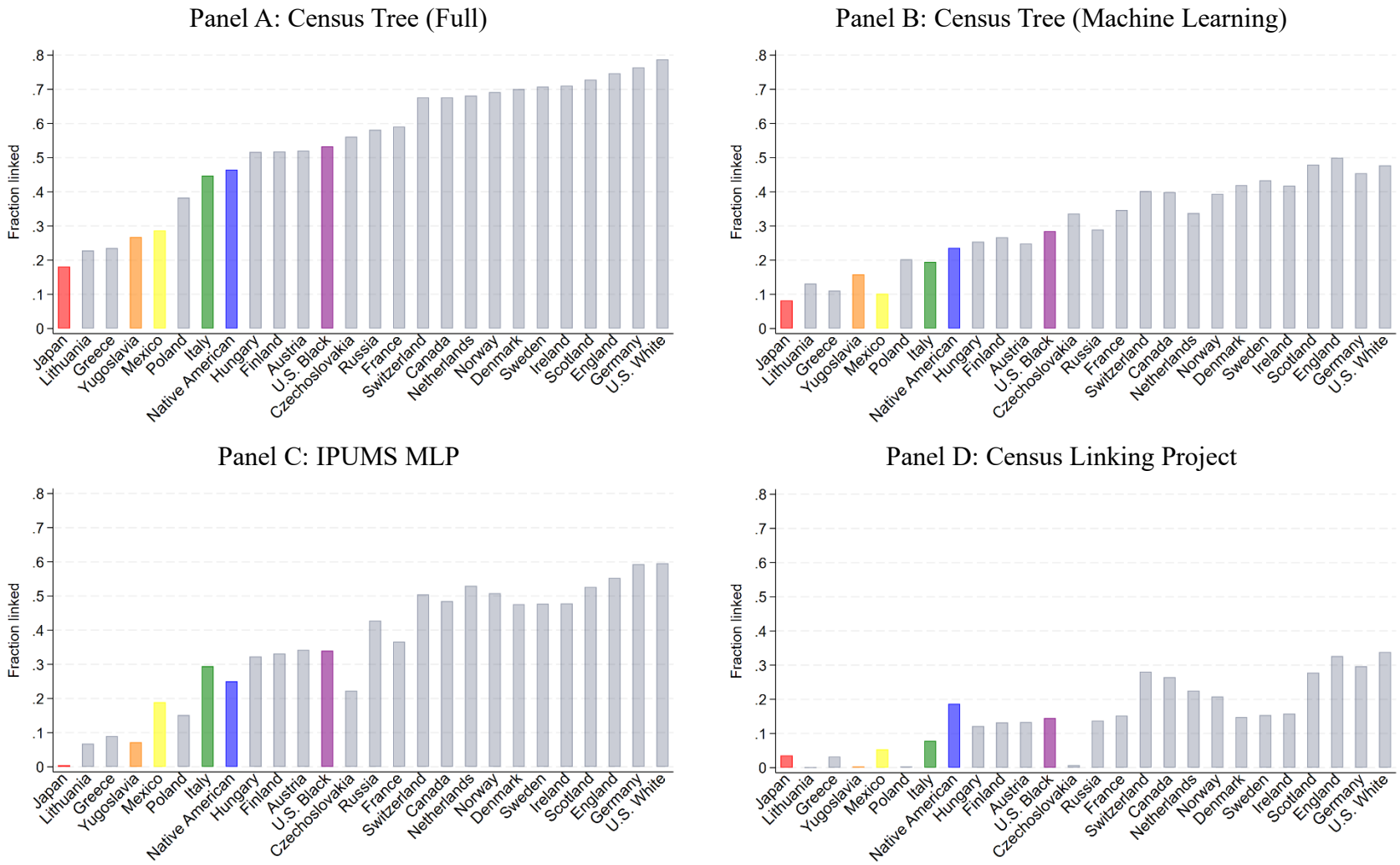
Michelman, Valerie, Joseph Price, and Seth D. Zimmerman. "Old boys' clubs and upward mobility among the educational elite." *The Quarterly Journal of Economics* 137, no. 2 (2022): 845-909.

Postel, Hannah M. "Record linkage for character-based surnames: evidence from Chinese exclusion." *Explorations in Economic History* 87 (2023): 101493.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. "Combining family history and machine learning to link historical records: The Census Tree data set." *Explorations in Economic History* 80 (2021): 101391.

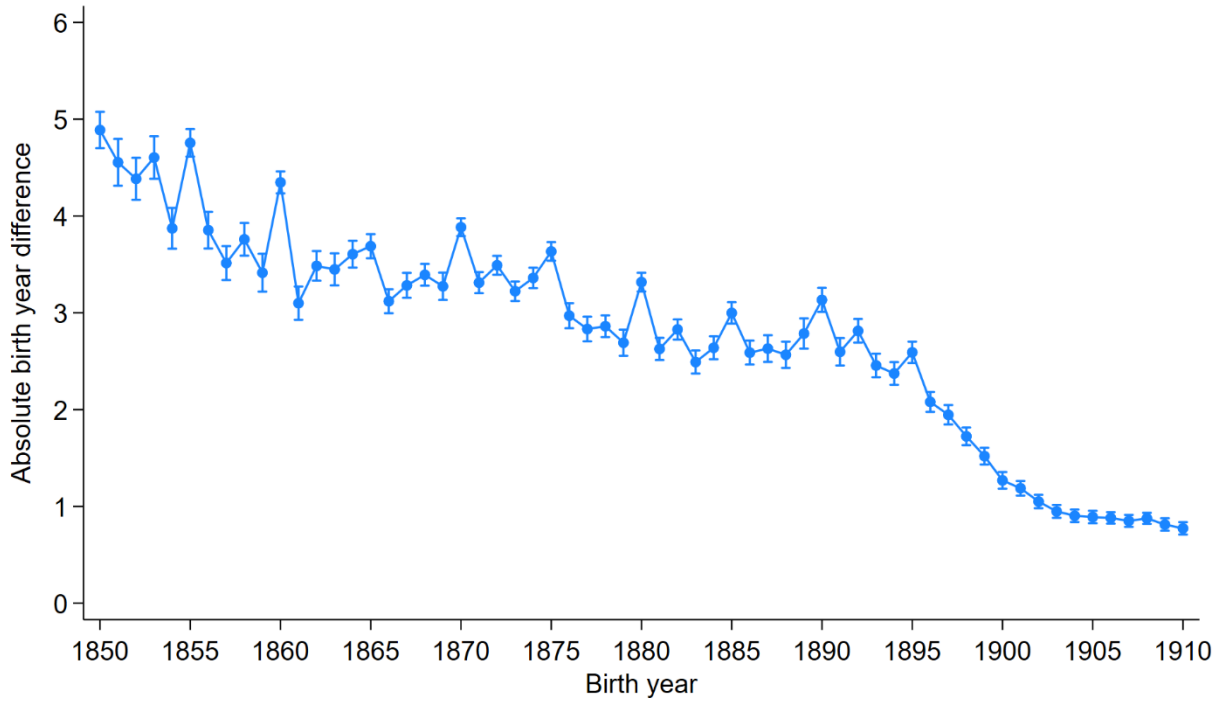
Ward, Zachary. "Intergenerational mobility in American history: Accounting for race and measurement error." *American Economic Review* 113, no. 12 (2023): 3213-3248.

Figure 1. Fraction of men in 1920 linked to 1910, by birth country and race



Notes: Denominators for each fraction are the number of men born in each country (or US-born who are Native American, Black, or White) and are 10 or older in 1920 and immigrated before 1910 (if applicable). The largest 23 immigrant groups are shown.

Figure 2. Birth Year Differences by Cohort among African Americans Linked in the Family Tree



Notes: Means and 95% confidence intervals of the absolute birth year difference for US-born Black men in the Family Tree, linked from 1910 to 1920. Birth year differences are shown by reported birth year in the 1920 census.

Table 1. Family Tree links in the Census Tree

Years	Total	U.S. Black	Native American	Italy	Mexico	Yugoslavia	Japan
1850 to 1860	4,553,096	12,797	403	97	189	10	0
1860 to 1870	5,228,303	19,908	557	187	399	37	2
1870 to 1880	8,024,987	106,483	471	504	1,101	7	5
1880 to 1900	11,859,990	202,412	2,071	1,560	2,418	7	16
1900 to 1910	28,874,030	1,585,126	13,096	33,625	8,713	53	170
1910 to 1920	30,873,542	721,606	36,554	82,159	11,208	4,296	654
1920 to 1930	27,199,151	336,912	45,971	100,268	16,186	13,408	2,057
1930 to 1940	27,448,446	345,476	36,110	131,851	24,528	15,732	3,645

Notes: Number of links from the Family Tree on FamilySearch.org which are included in the Census Tree. Categories include U.S.-born Black and Native American individuals, as well as immigrants born in Italy, Mexico, Yugoslavia, and Japan.

Table 2. Fraction of Family Tree links that pass the blocking criteria for the Census Tree or IPUMS MLP

	U.S. Black	Native American	Italy	Yugoslavia	Mexico	Japan
Census Tree – ML						
Sex	0.991	0.992	0.996	0.998	0.992	0.994
Birth year	0.846	0.889	0.889	0.890	0.769	0.901
Birthplace	0.970	0.945	0.983	0.978	0.818	0.969
First name	0.815	0.831	0.789	0.906	0.749	0.500
Last name	0.594	0.600	0.643	0.773	0.655	0.435
All	**0.471**	**0.492**	**0.553**	**0.666**	**0.404**	**0.379**
IPUMS MLP – Step I						
Sex	0.991	0.992	0.996	0.998	0.992	0.994
Birth year	0.846	0.889	0.889	0.890	0.769	0.901
Birthplace	0.970	0.945	0.973	0.013	0.818	0.969
First name	0.923	0.912	0.892	0.956	0.950	0.857
Last name	0.910	0.903	0.890	0.948	0.940	0.842
All	**0.738**	**0.753**	**0.767**	**0.012**	**0.587**	**0.730**

Notes: We evaluate Family Tree links between 1910 and 1920 for various racial and ethnic groups, as observed in the 1920 US census. Categories include U.S.-born Black and Native American individuals, as well as immigrants born in Italy, Mexico, Yugoslavia, and Japan. The Census Tree – Machine Learning algorithm evaluates candidate links that satisfy 5 blocking criteria: exact match on sex and birthplace across census years, an absolute birth year difference no greater than 3 years, and an exact match of standardized first and last names using the NYSIIS encoding. The IPUMS MLP – Step I algorithm evaluates candidate links that satisfy similar criteria: exact match on sex and birthplace across census years, an absolute birth year difference no greater than 3 years, and first and last names that each match on either the first letter or at least one bigram, as well as first and last names with a Jaro-Winkler string similarity of at least 0.7. Birthplaces and name strings are evaluated slightly differently: each algorithm uses a distinct crosswalk to standardize common name variations (e.g. Mike is matched to Michael) and the Census Tree – ML algorithm uses a broader definition of birthplaces, which allows Yugoslavia to match any European country.