Advancing the Frontier for Linking Subgroups in U.S. Census Records

Adrian Haws
Cornell University

Kasey Buckles
University of Notre Dame

Joseph Price
Brigham Young University

March 12, 2025

Abstract

Advances in automated record linkage for historical U.S. censuses have cultivated rich opportunities for empirical research in the social sciences. Although various linking methods demonstrate high match rates without sacrificing accuracy, concerns of population representativeness remain. In this paper, we show that several racial minority and immigrant subgroups of the U.S. population have match rates at less than half of the match rates for the U.S.-born white population, among both unsupervised and supervised linking algorithms. Large disparities in match rates suggest that re-weighting procedures may not be sufficient to overcome unobserved selection into linkage, threatening internal validity of research designs. In addition, low match rates further restrict sample sizes for studies focused on population subgroups. We demonstrate specific weaknesses in automated linking methods and how these weaknesses contribute to disparities in match rates. We then propose concrete solutions, and assess how these solutions affect representativeness and accuracy across subgroups. Our strategies are informed by a massive set of ground truth census links available through the Census Tree project.

JEL: C40, C81, N01, N31, N32

1 Introduction

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, currently available census linkages are not always suitable for studying subsets of the population, which are the focus of many research questions. Very low match rates for some groups raises two concerns. First, low match rates further restrict sample sizes in small groups, which can severely limit statistical inference. Second, large disparities in match rates across population subgroups raises further questions about unobserved selection into linkage, which can bias empirical estimates.

In this paper, we demonstrate that several racial minority and immigrant groups in the U.S. are linked across historical census records at less than half the rate of White U.S.-born men. This fact holds for three prominent linking approaches, which includes both unsupervised and supervised algorithms. We then provide evidence on the causes of subgroup disparities in match rates, by referencing a ground truth dataset that includes millions of hand-linked census records. This ground truth dataset, which was created by personal genealogists on the Family Tree at familysearch.org, is publicly available to researchers.

We also develop specific strategies for improving match rates for population subgroups, relying on the Family Tree links. These methods include creating a massive name standardization crosswalk, improving the flexiblity of blocking strategies, and including subgroups in training data. Preliminary results indicate that these strategies can greatly increase potential match rates among subgroups.

As an additional approach, we 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 match rates for under-represented groups, most notably for women who change their surnames when they marry (Althoff et al., 2025; Bailey et al., 2023; Buckles et al., 2024; Helgertz et al., 2022). Racial minority and immigrant groups have received 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. By taking a broader approach, we show that our proposed solutions can improve match rates across diverse population subgroups that are of interest to researchers.

Discussions of non-representativeness by race and immigrant origins may be less prominent in the historical linking literature because its causes are less obvious. The primary challenge in linking women is clear, but difficulties in linking Black Americans, for instance, is likely tied to a range of complex and interrelated social and economic issues. We contribute to the literature by illuminating key steps where automated linking algorithms are less favorable toward population subgroups. By using a data-driven approach to reveal these barriers, we also develop a general approach to overcoming them.

2 Subgroup Disparities in Matched Census Samples

2.1 Historical Census Linking Methods

We evaluate shortcomings in linking population subgroups among both supervised (using ground truth data) and unsupervised (no ground truth data) linking methods. Abramitzky et al. (2012, 2014) (henceforth ABE) formalize a set of rules similar to Ferrie (1996) to find unique matches on the basis of names, birthplaces, and birth years.[1] Some researchers prefer using rules-based linking methods because they are intuitive and inexpensive, not requiring the costs of labeling a ground truth data set. In our paper we compare the population representativeness of ABE – NYSIIS Conservative links (one of several variations of the ABE method) to the representativeness of supervised machine learning links, but the qualitative patterns we observe are similar across variations.

Supervised machine learning methods have also been applied to linking historical census records, including the IPUMS Multigenerational Longitudinal Panel (Helgertz et al., 2022) ()enceforth IPUMS MLP) and machine learning links from the Census Tree ((Price et al., 2021; Buckles et al., 2024).[2] These algorithms are designed to flexibly ”learn” from patterns in a labeled ground truth dataset. Supervised machine learning algorithms vary in their complexity, but are each designed to accurately predict when a candidate link is correct or incorrect. (Buckles et al., 2024) show that recent supervised machine learning methods obtain higher match rates than unsupervised methods, while also performing as well or better on accuracy in a random hand-checked sample.[3]

We compare links from the Census Tree, IPUMS MLP, and ABE because these are among the most widely used in recent research. Our primary emphasis is on the Census Tree – Machine Learning (Census Tree – ML) links. Because the Census Tree – ML links are a subset of the full

[1] Expectation Maximization (Fellegi and Sunter, 1969) has also been adapted for linking historical census records (Abramitzky et al., 2020) This approach falls under the category of unsupervised machine learning, rather than unsupervised rules-based methods.

[2] Feigenbaum (2016) developed an earlier machine learning approach to link men from the Iowa State Census to the U.S. Census.

[3] The authors find that machine learning links in the Census Tree have higher accuracy than the ABE – NYSIIS Standard method, but more conservative versions of the ABE method likely would perform better on accuracy (Abramitzky et al., 2021)

Census Tree (Price et al., 2021; Buckles et al., 2024), the subgroup match rates we report are lower than in the Census Tree.[4] A key difference between the Census Tree – ML links and the IPUMS MLP is in the ground truth data used to train the models. In contrast to the IPUMS MLP, which uses fewer than 1500 ground truth links, the Census Tree – ML algorithms are trained on over 100 million links. In Section 3 we discuss the characteristics of this massive set of training data and their implications for the resulting linked datasets.

Automated linking methods are not equipped to link women who change their surnames when they marry, unless additional information is available (e.g. marriage certificates). Although the Census Tree – ML and IPUMS MLP links include some women, our goal of comparing population-level match rates across three linking methods requires that we restrict the sample to men. However, the challenges and solutions for linking population subgroups are equally applicable in cases where it is possible to link women.

2.2 Match Rates for Racial Minorities and Immigrants

Figure 1 displays match rates among 12 population subgroups in the 1920 U.S. Census. Subgroups include white and Black U.S.-born men, Native Americans, and immigrants from 10 world regions.[5] We pay special attention to 7 subgroups that have comparatively low match rates and populations of no less than 100,000; these subgroups are highlighted in the figure. Match rates are computed among individuals in the 1920 U.S. Census who report that they were born and living in the U.S. by 1910.[6]

We find that each linking method shows considerable variation in match rates. Subgroup-specific match rates vary from 0.13 to 0.48 for Census Tree – ML links, from 0.05 to 0.58 for IPUMS MLP links, and 0.04 to 0.33 for ABE – NYSIIS Conservative links. These disparities are especially striking because each linking method is distinct in its algorithm, blocking strategies, training data, and de-duplication methods. Across each method, match rates are much higher for U.S.-born white men and immigrants from northern and western Europe and Canada, with very low match rates for immigrants from outside Europe.

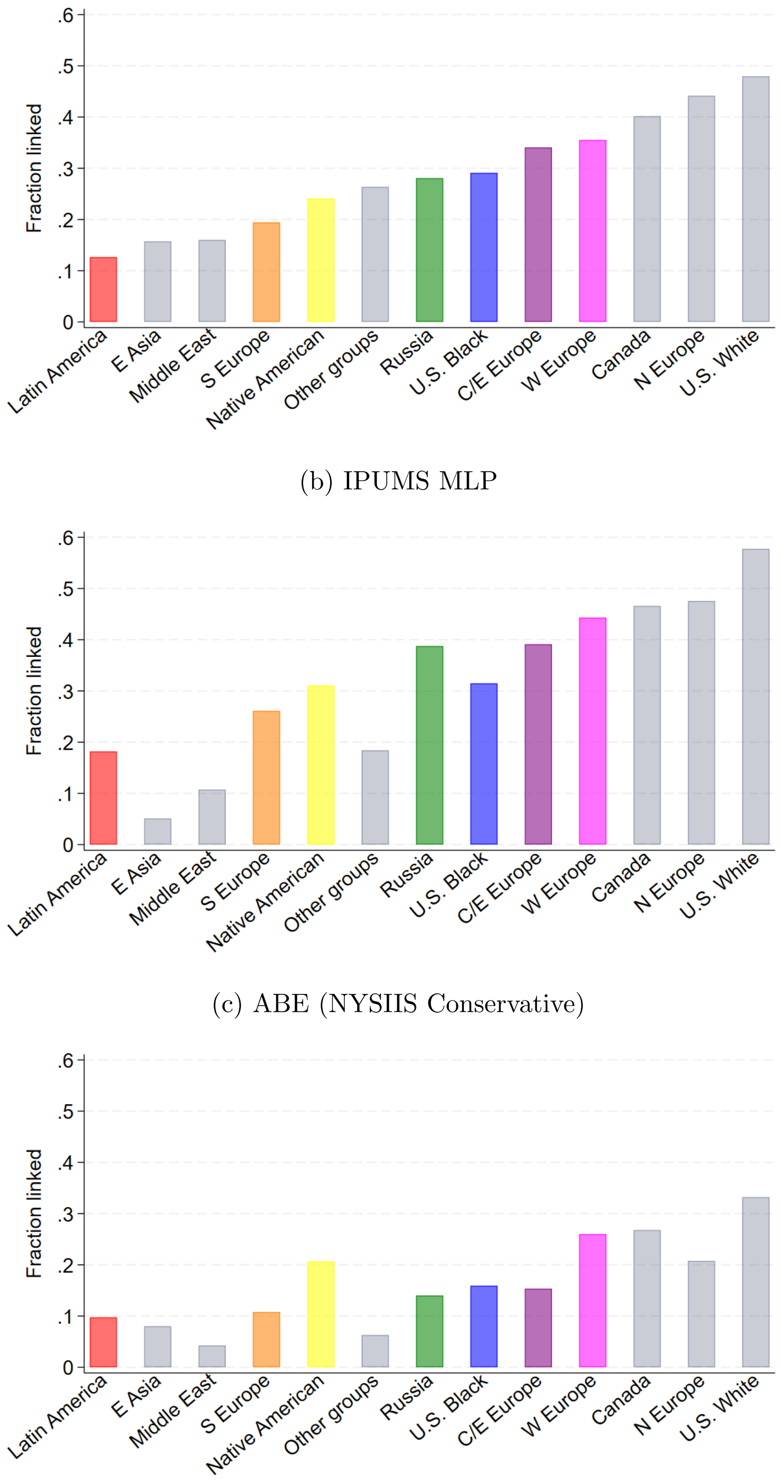
We additionally compute the ratio of each subgroup match rate to the match rate for U.S.-born white men. The lowest ratio for Census Tree – ML is for immigrants from Latin America, with 0.26, but the ratios for Southern Europe, East Asia, and the Middle East are all below

[4] We focus on this subset because it is the most flexible aspect of the combined linking process, but the full Census Tree includes links from the Family Tree, Census Tree – ML, additional machine learning links provided by the FamilySearch software, and automated links from the IPUMS MLP and the Census Linking Project (i.e. ABE links; see Abramitzky et al. (2021)). The Census Tree includes a process for adjudicating disagreements across sources when they occur.

[5] World regions are defined using categories of the "BPL" variable provided by IPUMS USA (Ruggles et al., 2024, 2025)

[6] This method of computing match rates relies on accurate information on ages and timing of immigration. As we show in Figure 2, children are much more likely to have accurate ages. Although there may be concerns of inaccurately reported immigration timing, historical accounts note that enumerators were instructed to emphasize accuracy in this field (Magnuson, 1995)

Figure 1: Fraction of men in 1920 linked to 1910, by birth region and race



(a) Census Tree (Machine Learning)

(b) IPUMS MLP

(c) ABE (NYSIIS Conservative)

0.5. That is, the match rates for four major subgroups are less than half the match rates for U.S.-born white men. The lowest ratio for IPUMS MLP is 0.09 (for East Asia), and the same subgroups have ratios below 0.5 (also including Latin America, Southern Europe, and the Middle East). ABE – NYSIIS Conservative has seven subgroups below 0.5, including a ratio of 0.48 for U.S.-born Black men. The lowest ratio for ABE – NYSIIS Conservative is 0.13 (for the Middle East). These comparisons show that representativeness across race and immigrant origins is a substantial weakness, even among linking methods that use concrete linking rules and only information that is stable across a person's lifetime. Through the rest of this paper we leverage a large set of ground truth links to demonstrate key drivers of these disparities, which are applicable across linking methods. We also demonstrate an improved linking approach that uses ground truth data that is publicly-available through the Census Tree.

3 The Family Tree Dataset

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 becomes more comprehensive and accurate over time.

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 machine learning algorithms. 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 (e.g. maiden names and aliases).

Although users on FamilySearch occasionally attach census records to the incorrect profile, the FamilySearch software helps users to find and fix these mistakes. In a blind 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 verified as correct. Other researchers have used the Family Tree links as "ground truth" to measure the quality of their automated census links (Bailey et al., 2023).

The Family Tree provides a much larger training set than would be feasible by hiring research assistants or professional genealogists. A key advantage of having such a large training set is that it includes links for very small subsets of the population.[7] Table 1 shows, for instance, that the Family Tree includes over 6,500 links between 1910 and 1920 census records for immigrant men from Latin America. Immigrants from Latin America comprised just 0.6 percent of the U.S. population in 1920. In contrast, the training data for IPUMS MLP links contains a total of 119 ground truth links for all immigrants.

Table 1: Family Tree Links, 1910–1920

Group	Men	Women
U.S. White	14,063,164	13,713,470
U.S. Black	357,728	350,037
Central and Eastern Europe	343,671	299,028
Russia	62,219	54,831
Western Europe	53,051	41,849
Southern Europe	47,706	37,380
Northern Europe	35,676	33,177
Native American	18,536	17,466
Canada	15,160	14,932
Latin America	6,516	5,922
Middle East	790	503
East Asia	598	387
Other Groups	15,393	13,937
Total	**15,477,938**	**15,012,614**

Similar to other supervised machine learning approaches, the training data available through the Family Tree is not selected randomly from the population. Table 1 shows, for instance, that 2.3 percent of Family Tree links between 1910 and 1920 are for Black men and 6.7 percent are for immigrants. The training data for IPUMS MLP contains 88 Black men (6.5 percent of the sample) and 119 immigrant men (8.8 percent of the sample). Both of these methods under-represent Black men—at 9.9 percent of the population in 1920, and immigrants—who comprise 13.3 percent of the 1920 population. Although it is clear that these linking methods are not less representative than unsupervised methods, it is possible that a more population-representative training set could result in improved representativeness of the linked sample. We explore this question in Section 5.

For the purposes of our paper, a primary benefit of the Family Tree links is that they allow us to diagnose weaknesses in linking population subgroups, as well as develop improved approaches to linking. The Family Tree links we use in this paper can be downloaded for free at censustree.org. The Census Tree, described by Price et al. (2021) and Buckles et al. (2024), includes links from the Family Tree as well as links from automated methods such as Census

[7] A massive training set also improves linking algorithms by including more variation in the types of links and allowing a high-dimensional set of feature comparisons.

Tree – ML.[8]

4 Challenges in Linking Subgroups

All linking methods reduce the set of comparisons by requiring agreement on a core set of variables. In a machine learning context this is referred to as "blocking", but a similar set of rules are used in linking methods such as ABE.[9] These rules are designed to limit candidate links to those with a reasonable likelihood of being correct, which is also important for computational feasibility. However, blocking sets an upper limit on match rates. We use the Family Tree links to show that fewer individuals in racial minority and immigrant subgroups make it past these initial barriers than U.S. white men.

We first look at birth year differences. All of the linking methods we discuss apply a rule that candidate links must not have a birth year difference greater than 3 years (in absolute value). It is quite common for individuals to report different ages across census years, for a variety of reasons. Household reporters to the enumerator might use rounding or guess at someone's age. Additionally, prior to 1900 there was less emphasis on knowing your birth date, and people who were illiterate may be less likely to consistently report their age. There was also misunderstanding of whether to report your age at the current time, or the age you would be at the end of the year.

Figure 2 shows that this rule severely limits the number of Family Tree links that would be possible to identify using automated linking methods. Except for children, as low as 60 percent of Family Tree links for Black Americans have a birth year difference within 3 years. In contrast, around 90 percent of white Americans have similar birth years. We also find that birth years are farther apart for earlier birth cohorts, and there appears to be a substantial amount of rounding to the nearest five years. Native Americans are similarly affected. The fraction of links that are within 5 years for Black Americans and Native Americans is similar to the fraction of links for white men that are within 1 year.

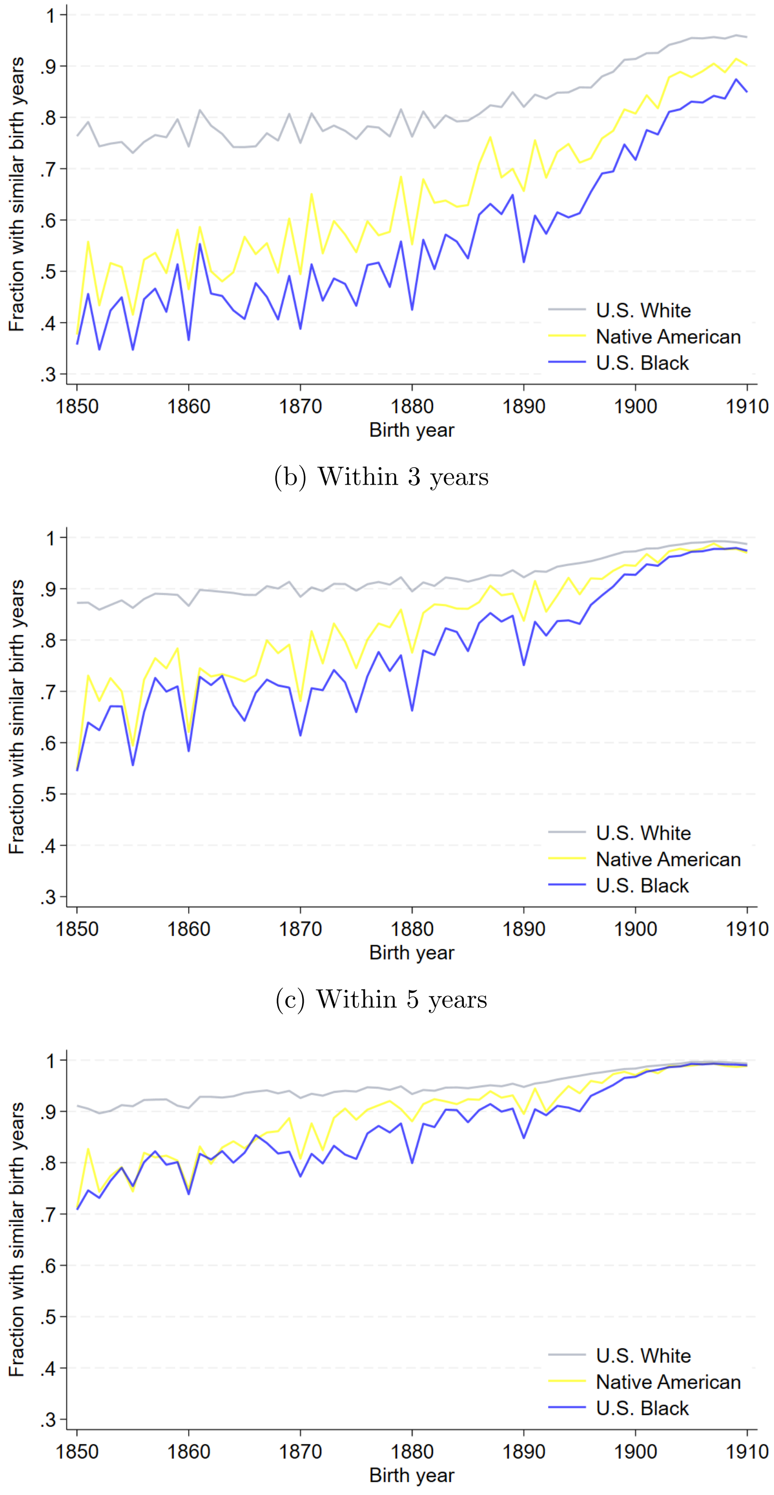
Additionally, all three linking methods require a match on birth place. The census records birth country for immigrants, but country boundaries often change. In these cases, there may be confusion of whether to report the name of the place where an individual was born or the current name of the place. World War I saw several notable boundary changes, including the creation of Poland, Czechoslovakia, Yugoslavia, and the Baltic nations. Figure 3 shows that very low fractions of Family Tree links with a 1910 reported birthplace in the former Austria-Hungary report the same country in 1920.

Another challenge in blocking is requiring similar names. Less than 60 percent of Family Tree links for immigrants from Southern Europe match on the NYSIIS encoding of first and last

[8] 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.

[9] See Price et al. (2021) for a discussion of blocking approaches.

Figure 2: Fraction of 1910–1920 Family Tree links with similar birth years

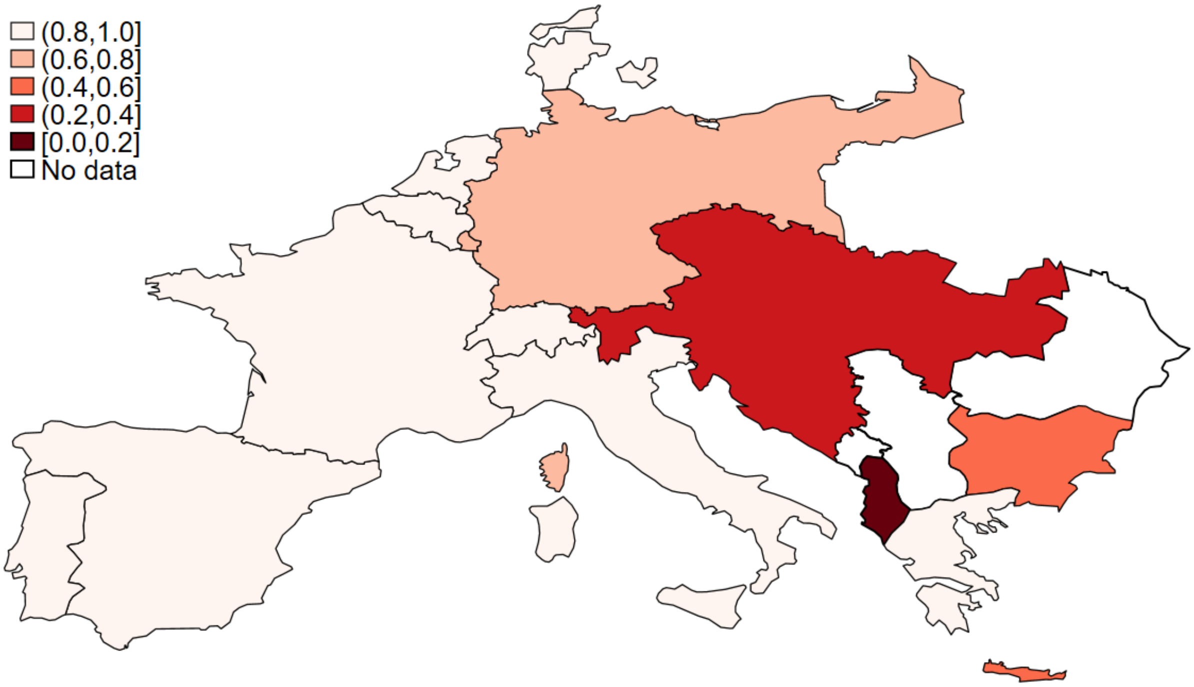


(a) Within 1 year

(b) Within 3 years

(c) Within 5 years

Figure 3: Fraction of 1910–1920 Family Tree links matching on birth country (defined in 1910)



names, after converting common nicknames to their standardized versions. This is the blocking requirement for Census Tree – ML and ABE – NYSIIS Conservative. In contrast, over 70 percent of U.S.-born white men match on names.

5 A Machine Learning Approach for Linking Key Subgroups

We develop targeted improvements to blocking, pre-processing, and the composition of training data, which all rely on the Family Tree data. Our strategies are especially compatible for machine learning. See Price et al. (2021) and Helgertz et al. (2022) for more discussion of machine learning applications to historical census linking. We focus on improving machine supervised learning algorithms, because these generate higher match rates and can fully leverage the training data available from the Family Tree. However, many of the strategies we describe can also be applied to unsupervised methods.

5.1 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 encoding, such as the NYSIIS algorithm, is the most common method for pre-processing names. This method can sometimes account for differences, but it can fail. For instance, Mike and Michael often refer to the same person, but yield different NYSIIS values.

To solve this problem, a common first step is to convert nicknames. ABE and Census Tree – ML use crosswalks that convert just over 2,000 first names to a standardized version. These cover the most common instances, but may not be effective for population subgroups that are more likely to have errors in the census data, driven by language and dialect differences. It is especially difficult to account for a wide variety of possible mis-spellings that can cause issues, including for surnames. For example, the surname ”De La Fuente” is often shortened to ”Fuente”.

It would be very difficult to analyze all possible versions of how Greek, Spanish, Italian names might be transliterated into English. We develop a data-driven approach to create first and last name crosswalks, using the full set of 158 million publicly-available Family Tree links for men, across all years, and 149 million links for women. This is a many-to-one mapping, converting into more commonly used names.

The first name crosswalk for men maps 114,000 unique spellings of first names to 10,000 standardized versions (after removing extraneous characters, first initials, and middle names). For women, we map 113,000 variations to 8,000 standardized names. Surnames matter as well, and these have not previously been standardized by linking methods. We map 1.9 million unique spellings of surnames to 0.5 million standardized versions. The crosswalk contains 36 different spellings of the Italian surname ”Bevilacqua”, which convert to 23 different NYSIIS values. Many of these would not be possible to link without the crosswalk. In Table 2 we compare the fraction of ground truth links that are represented by name standardization crosswalks from ABE, the previous Family Tree crosswalk, and the expanded crosswalks.[10]

5.2 Expanding the Set of Candidate Links

We adjust blocking strategies in a few key ways, as shown in Table 3. For immigrants, we allow candidate links to have a different country that is in the same region. This accounts for boundary changes. We allow more flexible surname matches in the (somewhat rare) case that the indexed name is indicated as uncertain (containing a ”?”, ”*”, or the word ”or” separating two names). Immigrants are 1.7x more likely than U.S.-born white men to have an uncertain last name. Finally, we adapt the birth years based on evidence shown in Figure 2.

[10] The crosswalks do not include names where the most common name is the same, such as Michael and Michael.

Table 2: Name Standardization Performance Measures

	ABE	Family Tree	Family Tree – Expanded	
	First names	First names	First names	Last names
Fraction of distinct names standardized				
U.S. Black	0.011	0.011	0.268	0.433
Native American	0.079	0.088	0.366	0.269
Latin America	0.021	0.021	0.172	0.251
Western Europe	0.069	0.065	0.401	0.314
Southern Europe	0.016	0.014	0.165	0.191
Central and Eastern Europe	0.023	0.022	0.266	0.244
Russia	0.038	0.035	0.295	0.216
Fraction of population with a standardized name				
U.S. Black	0.259	0.230	0.354	0.132
Native American	0.217	0.193	0.284	0.168
Latin America	0.166	0.212	0.226	0.220
Western Europe	0.243	0.145	0.208	0.230
Southern Europe	0.254	0.218	0.291	0.224
Central and Eastern Europe	0.259	0.172	0.255	0.205
Russia	0.304	0.194	0.280	0.189

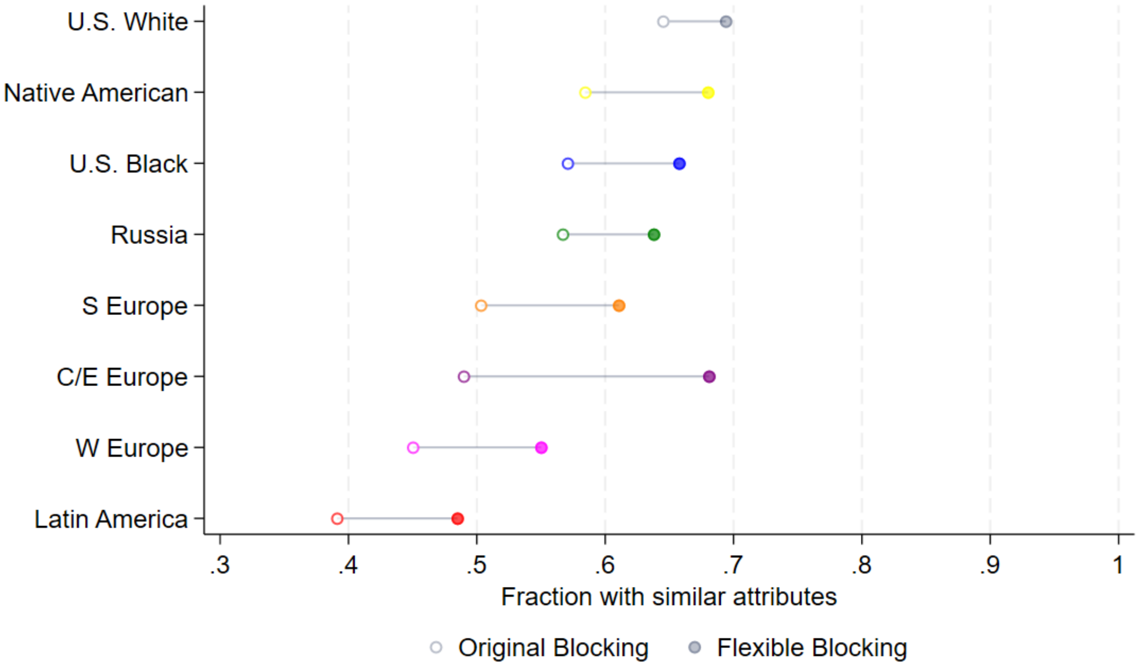
Table 3: Blocking Rules

Match criteria	**Subset of population**
Sex	All
Birth state	U.S.-born
Birth region	Immigrant
NYSIIS of standardized first name	All
NYSIIS of standardized last name	All
At least one adjusted bigram match of standardized last name and Jaro-Winkler similarity score > 0.7	Uncertain last name
Birth year within 1 year	Born after 1900
Birth year within 2 years	Born after 1900 and Black, Native American, or immigrant from Latin America
Birth year within 3 years	Born before 1901
Birth year within 5 years	Illiterate; or born before 1901 and Black, Native American, or immigrant from Latin America
Birth year within 5 years or exactly 10 years apart	Birth year ending in 0 or 5 and conditions for birth year within 5 years

6 Results

Figure 4 shows that our blocking strategies raise the upper bound on match rates, according to subgroup-level ground truth links. The fraction of Family Tree links that pass the more flexible blocking strategies increases much more for all of the population subgroups that we focus on than for U.S.-born white men. Thus, these blocking strategies likely have the dual benefit of increasing match rates and improving representativeness. At this point our results are very preliminary, but we will later provide statistics on final match rates and the accuracy of these matches.

Figure 4: Fraction of 1910–1920 Family Tree links with similar characteristics



7 Automated and Manual Methods for Near-Complete Linkages in Small Subgroups

We have shown that match rates for population subgroups are often quite low, which introduces selection bias and limits statistical inference. Some research projects are focused exclusively on small subgroups of the population, which could present an opportunity to overcome these challenges. Although 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.

8 Conclusion

Recent advances in automated census linking methods have made it possible to answer many empirical questions in economics, sociology, political science, and other fields. Automated methods demonstrate that it is possible to obtain large linked samples using historical census data, while maintaining high accuracy. There has been comparatively less emphasis, however, on the ability of these methods to link minority groups. In many research applications, estimates can be biased if subgroups of the population are systematically underrepresented among linked census samples. For example, Ward (2023) demonstrates that intergenerational mobility in the U.S. has previously been overestimated because linked samples did not include Black Americans. In some applications, under-representativeness of minority groups can also limit statistical inference.

This paper highlights the role of publicly-available genealogical data in improving match rates for population subgroups. We use millions of links created by users of a collaborative Family Tree to expand the set of census links for several racial minority and immigrant subgroups. The Family Tree provides a diverse set of ground truth census links which informs pre-processing, blocking, and supervised machine learning models. We also describe an approach for combining automated record linking with manual methods to attain very high match rates for small population subgroups. These strategies demonstrate the potential to increase the representativeness of linked samples by increasing previously low match rates for population subgroups. As an additional benefit the total number of census links in the population also increases.

References

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review*, 102(5):1832–1856, May 2012. doi: 10.1257/aer.102.5.1832.

Ran Abramitzky, 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(3):467–506, 2014. doi: 10.1086/675805.

Ran Abramitzky, Roy Mill, and Santiago Pérez. Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):94–111, 2020. doi: 10.1080/01615440.2018.1543034.

Ran Abramitzky, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. Automated Linking of Historical Data. *Journal of Economic Literature*, 59(3):865–918, 2021. doi: 10.1257/jel.20201599.

Anna Aizer, 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(2):492–529, 2024a. doi: 10.1257/app.20210816.

Anna Aizer, 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*, 139(4):2579–2635, November 2024b. doi: 10.1093/qje/qjae016.

Lukas Althoff, Harriet Brookes Gray, and Hugo Reichardt. America's Rise in Human Capital Mobility. 2025.

Martha Bailey, 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(3):138–159, 2023. doi: 10.1080/01615440.2023.2239699.

Kasey Buckles, Adrian Haws, Joseph Price, and Haley E.B. Wilbert. Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project, 2024.

Gregory Clark. The inheritance of social status: England, 1600 to 2022. *Proceedings of the National Academy of Sciences*, 120(27):e2300926120, 2023. doi: 10.1073/pnas.2300926120.

Dora L. Costa, 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(2):79–95, April 2017. doi: 10.1080/01615440.2016.1250022.

James Feigenbaum. Automated census record linking: a machine learning approach, 2016.

Ivan P. Fellegi and Alan B. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

Joseph P. Ferrie. A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, October 1996.

Jonas Helgertz, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles, and Catherine A. Fitch. A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55(1):12–29, January 2022. doi: 10.1080/01615440.2021.1985027.

Diana Lynn Magnuson. *The making of a modern census: The United States census population, 1790-1940.* Ph.D., University of Minnesota, United States – Minnesota, 1995.

Valerie Michelman, Joseph Price, and Seth D Zimmerman. Old Boys' Clubs and Upward Mobility Among the Educational Elite*. *The Quarterly Journal of Economics*, 137(2):845–909, May 2022. doi: 10.1093/qje/qjab047.

Hannah M. Postel. Record linkage for character-based surnames: Evidence from chinese exclusion. *Explorations in Economic History*, 87:101493, January 2023. doi: 10.1016/j.eeh.2022.101493.

Joseph Price, 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*, 80:101391, April 2021. doi: 10.1016/j.eeh.2021.101391.

Steven Ruggles, Matt A. Nelson, Matthew Sobek, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Evan Roberts, and J. Robert Warren. IPUMS ancestry full count data: Version 4.0 [dataset]. Minneapolis, MN: IPUMS, 2024.

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, Renae Rodgers, Jonathan Schroeder, and Kari C.W. Williams. IPUMS USA: Version 16.0 [dataset]. Minneapolis, MN: IPUMS, 2025.

Zachary Ward. Intergenerational Mobility in American History: Accounting for Race and Measurement Error. *American Economic Review*, 113(12):3213–3248, 2023. doi: 10.1257/aer.20200292.