NBER WORKING PAPER SERIES

BREAKTHROUGHS IN HISTORICAL RECORD LINKING USING GENEALOGY DATA: THE CENSUS TREE PROJECT

Kasey Buckles
Adrian Haws
Joseph Price
Haley E.B. Wilbert

Working Paper 31671
http://www.nber.org/papers/w31671

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023, Revised November 2024

This work is supported by grants from the Russell Sage Foundation (G-1063) and the National Science Foundation (SES-2049762). We are grateful to the excellent research assistants at the University of Notre Dame and the BYU Record Linking Lab for their work on this project, including Brianna Felegi, Nate McGhie, and Tommy Morgan. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2023 by Kasey Buckles, Adrian Haws, Joseph Price, and Haley E.B. Wilbert. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project
Kasey Buckles, Adrian Haws, Joseph Price, and Haley E.B. Wilbert
NBER Working Paper No. 31671
September 2023, Revised November 2024
JEL No. C81, J10, N01

ABSTRACT

The Census Tree is the largest-ever database of record links among the historical U.S. censuses, with over 700 million links for people living in the United States between 1850 and 1940. These high-quality links allow researchers in the social sciences and other disciplines to construct a longitudinal dataset that is highly representative of the population. In this paper, we describe our process for creating the Census Tree, beginning with a collection of links contributed by the users of a free online genealogy platform. We then use these links as training data for a machine learning algorithm to make new matches and incorporate other recent efforts to link the historical U.S. censuses. Finally, we introduce a procedure for filtering the links and adjudicating disagreements. Our complete Census Tree achieves match rates between adjacent censuses that are between 69 and 86% for men, and between 58 and 79% for women. To demonstrate the advantages of the Census Tree, we extend the work of Abramitzky, Boustan, Jácome, and Pérez (2021) to include intergenerational mobility estimates for additional immigrant nationalities and for women.

Kasey Buckles
University of Notre Dame
Department of Economics
3052 Jenkins-Nanovic Halls
Notre Dame, IN 46556
and NBER
kbuckles@nd.edu

Joseph Price
Department of Economics
Brigham Young University
2129 WVB
Provo, UT 84602
and NBER
joseph_price@byu.edu

Adrian Haws
Cornell University
434 Warren Hall
Ithaca, NY 14853
aah226@cornell.edu

Haley E.B. Wilbert
University of Notre Dame
Department of Economics
3052 Jenkins-Nanovic Halls
Notre Dame, IN 46556
hwilbert@nd.edu

For additional information about the Census Tree and to access data and code, please see censustree.org.

I. Introduction

Record linking, or the process of combining a subject's information from multiple datasets, is often necessary for empirical work in history, medicine, and the social sciences. These links allow the researcher to observe a person over time, to study relationships among variables that are not available in a single data source, and to identify connections between people in families and communities. Recent advances in record linking have been facilitated by growing access to restricted-use data that include stable and unique personal identifiers (e.g. social security numbers, registry numbers, or exact birth dates) that can be used to determine that two records correspond to the same person (Chetty et al. 2014; Chetty, Hendren, and Katz 2016; Mazumder 2005; Black, Devereux, and Salvanes 2005; Kleven, Landais, and Søgaard 2019). Unfortunately, many datasets that researchers would like to link—including many historical or publicly available sources—do not include these identifiers. In this situation, researchers must try to find unique matches using relatively stable characteristics like names, birth years, and birth places. These requirements can result in non-representative samples; in particular, women have been omitted entirely from several notable linking efforts because their surnames typically change when they marry (e.g. Fogel and Wimmer 1992; Feigenbaum 2018; Abramitzky et al. 2020; Collins and Wanamaker 2022).

In the Census Tree project, we use information provided by members of the largest genealogy research community in the world to create hundreds of millions of new links among the historical U.S. Censuses (1850-1940). The users of the platform link data sources—including decennial census records—to the profiles of deceased people as part of their own family history research. In doing so, they rely on private information like maiden names, family members' names, and geographic moves to make links that a researcher would never be able to make using the observable information. To date, users have created over 317 million census-to-census pairs, nearly half of which are for women.

We describe our process for adding to these links using a machine learning model that employs the user-created links as training data. We also add pairs identified by other recent linking methods and develop a process to verify the quality of the matches and to adjudicate disagreements between methods. The result is the publicly-available Census Tree dataset, which contains over 700 million links among the 1850-1940 censuses.[1] The data include an unprecedented number of links for women (314 million) and Black Americans (41.5 million). We show that the Census Tree links

[1] All of the data described in this paper are available at censustree.org, along with the code and training data for the machine learning methods and the code for creating the full Census Tree.

are high quality and yield samples that are highly representative of the population. We also discuss features of the data that allow researchers to make tradeoffs between higher match rates (recall) and more accurate matches (precision), and to create samples that are representative of a target population.

The Census Tree will enable new work in history and the social sciences on topics including the intergenerational transmission of socioeconomic status, assortative mating, and the long-term effects of public policies, events, family structure, communities, and the early childhood environment.[2] To demonstrate the potential of the Census Tree, we revisit the work of Abramitzky, Boustan, Jácome, and Pérez (2021)—henceforth ABJP—who show that the children of immigrants were more upwardly mobile on average than the children of the U.S.-born in the late 19th and early 20th centuries. We replicate this result using the Census Tree, and are able to increase the precision of estimates for each sending country. Furthermore, the Census Tree includes sufficient numbers of links to produce estimates for an additional ten countries, including countries from Central America and the Caribbean. We find that the sons of low-income immigrants from Mexico had significantly worse outcomes on average than sons of fathers from other countries, including U.S.-born Whites. We further extend ABJP by analyzing the mobility of women in a historical sample, and compare these to historical estimates for men and to the authors' modern estimates for women. While the patterns for daughters and sons are broadly similar, differences in marriage patterns contribute to gender gaps in mobility for some countries.

In the next section, we describe the online platform that is the source of our genealogy data, and the census-to-census links that the platform's users create. In Section III, we outline the process for adding links using our machine learning algorithm and from other sources to build the full Census Tree. We present results that summarize the match rates, precision, and representativeness of the Census Tree in Section IV, and introduce our replication of ABJP in Section V.

II. Genealogy Research on FamilySearch

A. The Platform

Founded in 1894, FamilySearch is "a nonprofit family history organization dedicated to connecting families across generations" (FamilySearch 2023a). Sponsored by the Church of Jesus

[22] See Appendix B for a bibliography of work that is already using the Census Tree to study these topics and more.

Christ of Latter-day Saints, FamilySearch introduced a free website featuring family history tools and digitized records in 1999. It has since become one of the most widely-used genealogy websites in the world, with over 400,000 visitors per day from 238 countries. The website also includes over two billion indexed historical records and over one billion unique individual profiles for deceased persons (FamilySearch 2023b).

FamilySearch.org has several features that contribute to its popularity among the genealogy community, including its sophisticated search tools, its enormous set of digitized and indexed historical records, and the fact that it is free to all users. But perhaps its most distinctive feature is that, rather than each user building their own tree, all users contribute to a single, interconnected family tree. The tree operates as a "wiki," in which users can edit and build on the contributions of others. As a result, FamilySearch users have collaborated to produce an incredibly comprehensive and accurate population-wide family tree that includes 1.55 billion people.

Critically for our purposes, users can attach digitized historical records to the profiles of people on the tree, including the decennial U.S. censuses from 1850 to 1940.[3] In cases where records in two different decennial censuses are linked to the same profile, this creates a user-made link that identifies the records as describing the same person. Thus, the process of record linking is "crowd-sourced" to millions of users with private information that helps them make links—including some information that would be unavailable to trained research assistants or machine learning algorithms. For example, family members often know their female ancestors' maiden names, which allows them to attach both childhood and adult records to the profile, thereby creating links for women that would not be possible using linking methods that rely on a name match. Users may also know details that make it possible for them to solve the problem of common names—they may know the names of other family members within the same household that allow them to correctly identify which "John Brown" is the right one among many choices. This information can also help them to confirm that two records are a match, even if the digitized spelling of the name is different or if other information is not an exact match.[4]

[3] The Census Bureau releases the full-count censuses to the public after 72 years. The 1950 census was released in April of 2022; our work to add it to the Census Tree will begin after a large number of FamilySearch users have linked these records to profiles. The 1890 census is not included in the set of historical decennial censuses as most records for that year were destroyed in a fire in 1921.

[4] Appendix Figure 1 shows the sources linked to "Delilah A. 'Minnie' Jenkins," who appears in the digitized censuses as Delila A Jenkins (1870), Deliah M Jinkins (1880), Minnie Sharone (1900), Minnie Shearom (1910), and Minnie Sherman (1920). The consistent presence of other family members across these records helps to confirm that they do reference the same person.

B. User-Made Links: The Family Tree

The set of over 317 million unique user-made links among the 1850-1940 censuses constitutes a dataset that we call the "Family Tree." The Family Tree contains between 24 and 48% of the possible matches between adjacent censuses for men, and the match rates for women are nearly as high.[5] How reliable are the Family Tree links, given that they are crowdsourced and not directly validated? To investigate this, we conduct an exercise in which we randomly selected 1,000 people in the 1910 census for whom it should be possible to find a match to 1900, as they are at least 10 years old and did not immigrate within the last ten years. Among these, we have a Census Tree link for 759 people, and a Family Tree link for 440. We then ask trained research assistants to use the full set of information available in each census record to classify each link as correct, incorrect, or unsure.[6] Among the Family Tree links, 98% were determined to be correct—an exceptionally high number that is consistent with a similar check conducted on a different sample in Price et al. (2021).[7] We describe the results of this exercise for the full Census Tree in Section IV.B.

One potential limitation of the Family Tree data is that the users may be a selected group. Among other possible factors, they have a demonstrated interest in family history and are able to access and use the internet. We explore this in Section IV, where we compare the observable characteristics of people who can be linked in the Family Tree and other datasets to the linkable population.

---

[5] Throughout the paper, the match rate is calculated as the number of people for whom a link is made to the previous census, divided by the number of people who are old enough to have been alive at the time of the previous census, with an adjustment for rates of immigration and under-enumeration. See Price et al. (2021) for a detailed description of how these match rates are calculated.

[6] Research assistants were instructed to code a match as "correct" if all observable information (birth year, gender, race, family members, location) matched, as "unsure" if most but not all information matched, and as "incorrect" if the records were very unlikely to refer to the same person.

[7] There is also external evidence that the user-provided information is high quality. Using data from a similar genealogy platform, Kaplanis et al. (2018) compare DNA data to information provided by the site's users, and conclude "that millions of genealogists can collaborate in order to produce high quality population-scale family trees" (p. 172). Furthermore, the creators of other linked datasets have used the Family Tree as a benchmark for measuring the quality of their own matches (Bailey et al. 2020), referring to genealogy data as the "gold standard" (Abramitzky et al. 2021a, 868).

III. Creating the Census Tree Dataset

Figure 1 illustrates the process we use to create the Census Tree dataset. We first generate links using our machine learning process, where we use Family Tree links to inform pre-processing and blocking and as training data for the model. We then include links obtained from other recent linking efforts and develop a process for filtering low-quality links and adjudicating disagreements. We elaborate on these steps in this section.

A. Machine Learning Using Training Data from the Family Tree

1. Pre-processing and blocking

We begin by preparing the data to be linked by the machine learning process, drawing on information provided by the user-made links. We standardize the names of places (states and countries) to correct misspellings and abbreviations. For names, we convert nicknames to a standard set of formal names, using a list of the most common nickname-name pairs observed in the Family Tree.

The computational costs of our machine learning process also require that we limit the set of potential matches by grouping the data into blocks based on features like name, birthplace, and birth year. A challenge when choosing the features to create the blocks is that the most stable features, like race, sex, or birth state, are not very unique. Requiring that the potential matches also have, for example, the same birth year, might exclude many true matches. We are able to test several blocking strategies to see how they perform when trying to recreate the links in the Family Tree data. Appendix Table 1 identifies in bold the variables that we use in our blocking strategy.

2. Training Data

We use millions of the user-created links from the Family Tree to train our machine learning models. After removing any non-unique links, we use the "true" links from the Family Tree to create a set of "false" links by identifying all other potential matches that satisfy the same blocking criteria but are not the same as the "true" link. For each of the 36 year-to-year pairs, we train the model using training data from those specific years; see Appendix Table 2 for the number of "true" links for each pair of years. The large size of our training data helps to improve the accuracy and number of record links (Feigenbaum, 2016; Gross and Mueller-Smith 2021). The size also ensures that we have sufficient support in the data for training the algorithm to make matches for under-represented groups, including thousands of observations for both women and Black Americans in the training data for all pairs of censuses that are 30 or fewer years apart.

Each census record contains basic information about the person's name, birth year, residence, demographic characteristics, household relationships, and occupation. To prepare the training data, we convert these variables into "features" that capture the rich amount of information available.[8] For example, when comparing the birth year between two records, we create four features: a binary variable indicating that the absolute difference between them is less than or equal to 3, a variable that is equal to the absolute difference in birth years, an indicator that the sign of the birth year difference is positive, and a measure of the age in the earlier census. Appendix Table 1 shows the full list of 70 features created across the nine censuses, and the years that the feature is available.[9]

3. Tuning the Model and Filtering Predictions

The supervised machine learning algorithm, XGBoost, uses gradient-boosted decision trees to assign a score to each potential link.[10] This score, between zero and one, is similar to a predicted probability of a link being "true" that could be calculated using a logistic regression. We use a cross-validation process with the training data to select the values of our model parameters—maximum tree depth and number of estimators—to optimize the model's performance. For each set of census years, we randomly select two-thirds of 500,000 training pairs to train a model and use the remaining third to test the out-of-sample performance. The model with the highest F1 score (balancing precision and recall) is then used with the full set of training data to produce the final model.[11] We provide the trained models for all 36 year-to-year pairs at censustree.org.[12]

As a way of getting "under the hood" of the machine learning algorithm, we calculate the importance of each of the 70 features used in the process of linking the 1900 and 1910 censuses, after the core set of features used for blocking. The importance measure is calculated as the average increase in accuracy across nodes of the decision tree which use the feature; this is the "gain" method of feature importance calculated by the XGBoost algorithm. Table 1 lists the fifteen most

[8] Names are not available in the publicly released versions of the IPUMS census files, but users can apply for the restricted-use versions of the data that include them.
[9] This extensive set of features benefits from indexed census variables provided by Ruggles et al. (2021) as well as geographic coordinates from the Census Place Project (Berkes, Karger, and Nencka 2023).
[10] Supervised methods require training data; unsupervised methods can be automated but do not require training data. We use the XGBClassifier package within the xgboost library in Python.
[11] The F1 score is calculated as (2PR)/(P+R), where P is precision and R is recall.
[12] The website also includes the full training dataset for 1900-1910.

important features for 1900-1910. The most important individual feature is the distance in miles between the towns of residence in the two censuses. This illustrates the value of the machine learning approach—using a traditional blocking and matching procedure, one would not want to require that two records be from the same (or nearby) towns, as people frequently move. However, if the person *is* living close to their location in the earlier census, that increases the probability that the records are a match. [13] Most of the other important features are variations on the characteristics most commonly used in blocking—birth year, name, and birth place.[14]

The machine learning algorithm generates a match score for each potential match within the blocking cell. We identify a pair of records as a match if it satisfies three conditions. First, it should have the highest match score among possible links. Second, it should have the highest sheet count, where the sheet count is the total number of individual links between the census pages containing the records. If record A and four additional records are linked to the sheet containing record B, then A and B have a sheet count of five; a high sheet count suggests that the same set of neighbors appears in both censuses, increasing our confidence that the match is true. [15] Third, we remove any remaining conflicts between the two years.[16] We tested this method using a "truth set" from the Family Tree and determined that over 98% of true links satisfy these conditions. We additionally remove a small set of links for women with consistent surnames but who transitioned from single to married between the census years we attempt to link.[17] This represents only 0.9% for 1900-1910 links for women because these cases are already penalized by the machine learning model.

---

[13] See Folkman, Furner, and Pearson (2018) and Price et al. (2021) for a more in-depth discussion of this issue, and for a demonstration of the effects of excluding geographic information from the set of features.

[14] See Appendix Table 3 for a ranking of the importance of feature categories for all adjacent census pairs.

[15] We calculate sheet counts using the set of potential links with a match score above 0.1. While many of these potential links are later removed from the sample, this match score criterion removes 92.5% of the blocked pairs between 1900 and 1910. The occurrence of multiple links between a set of sheets could almost never occur by random chance, as there are 40 million potential links remaining and about 2.8 trillion possible combinations of census sheets between 1900 and 1910.

[16] Conflicting links may persist with an exact tie for both the match score and sheet count. A more common case is where the highest match score for the 1900 record A occurs with the 1910 record B, but the highest match score for the 1910 record C occurs with record A. This second type of conflict would only persist if the A to B link also has the same sheet count as A to C.

[17] Because marital status is not available for the 1850 through 1870 censuses, we remove links for women who are married to the household head in the later census but have a different household relationship in the earlier census. This alternative strategy removes 4.6% of 1870-1880 links for women.

B. Additional Sources for Links

In addition to links from the Family Tree and the XG Boost algorithm, we incorporate links from the following sources.

1. *Census Linking Project*

The Census Linking Project (CLP) was the first effort to fully link the 1850-1940 decennial U.S. censuses and to make the links publicly available to the research community (Abramitzky et al. 2020). These links are based on traditional, unsupervised blocking and matching strategies that rely on names, birth dates, and birth places; see Abramitzky, Boustan, and Eriksson et al. (2021) for a detailed description of their process. The CLP data contain multiple sets of links, which use slightly different features and more or less conservative rules to identify matches. We use the NYSIIS Standard links, which use the New York State Identification and Intelligence System Phonetic Code to standardize names based on their pronunciation and require that the names be unique within the birth year. We choose this set because it has a high match rate, allowing us to include more links; we discuss this choice further below.

2. *Multigenerational Longitudinal Panel*

Helgertz et al. (2023) created the IPUMS Multigenerational Longitudinal Panel (MLP) of links between adjacent censuses.[18] They introduce an innovative two-step approach, in which they first use machine learning to obtain high-quality matches for men, and then link together other individuals in the same households in those two linked records.[19] This strategy allows them to match women as well as men.

3. *FamilySearch Hints*

FamilySearch has a proprietary machine-learning algorithm for identifying possible record links. They have provided us with two sets of these "hints" for U.S. census records. The first type of

[18] This was version 1.1 of the MLP data; in April 2024, IPUMS USA released MLP version 1.2 which includes links for 20- and 30-year intervals. Because the current public version of the Census Tree was constructed using MLP version 1.1, we refer to that version throughout the paper. On average, the Census Tree contains 2.38 times as many 20-year links and 2.89 times as many 30-year links as the MLP version 1.2. Future versions of the Census Tree will incorporate these new MLP links.

[19] The MLP household-based strategy is similar to the dyad and household matching methods that were part of the process described in Price et al. (2021). Because the MLP data contain nearly all of the additional links generated by these methods, we do not implement them here.

hints, which we call "profile hints," suggest to users that a census record might belong to a profile in their family tree. When census records from two different years are both "hinted" to the same profile, this creates a possible census link. The second type, which we call "direct hints," identifies a possible link directly between two census records. We have developed several tools that allow volunteers to validate both types of hints by attaching records to profiles on the Family Tree. In this way, these hints help to expand the set of user-made links on the Family Tree.

FamilySearch hints include many links for women, which is made possible by the large corpus of digitized records on the website (including marriage records) and by personal information available on person profiles (including dates of marriage and spouses' surnames). While we do not have access to FamilySearch's machine learning models, the methods employed by genealogy companies can be quite rich (Folkman, Furner, and Pearson 2018). We use match scores provided by FamilySearch to apply the same three-step filtering process described for our XGBoost model. As Table 2 shows, there are nearly 27 million FamilySearch "direct hints" that are part of the 1900-1910 Census Tree links, of which 0.5 million are not found by one of the other methods in the full linking process. A similar number of "profile hints" are used in our links.

C. Preparing the Data

1. *Filtering and Adjudication*

We combine unique links from the six sources described above: the Family Tree, XGBoost, the Census Linking Project, the Multigenerational Longitudinal Panel, FamilySearch profile hints, and FamilySearch direct hints. Because these various links may disagree, we filter them using the same sheet checking procedure described above. In this case, we calculate sheet counts using links from all six methods (without double-counting the same link from multiple methods), keep potential links which have the highest sheet count for each year, and drop any links with remaining conflicts.

2. *Creating Implied Links*

This step takes advantage of the fact that if records from two different censuses are linked to a record in a third census, the original two should also be a match. For example, if a link has been established between a person's 1900 and 1910 census records, and the 1910 record is linked to a 1920 census record, we can also link the 1900 record to 1920. As with other link sources, implied links are filtered by keeping potential links which have the highest sheet count for each year and dropping any links with remaining conflicts. We also remove implied links with an absolute birth year difference greater than three years. These implied links constitute our seventh and final source of record links.

3. *Creating the Crosswalks*

After creating the implied links, we conduct one final round of sheet-checking and drop remaining conflicts. We also add flags to identify the linking method(s) used to create each link. As we discuss below, the link source flags should be helpful in the event that a researcher wishes to exclude links made by a particular method. Table 2 shows the number of total and unique links provided by each of the different linking methods for 1900-1910, while Table 3 shows the number of sources that identify the links. Of the 47.4 million links between these two censuses, 7.9 million are identified by just one source, while 25.6 million are identified by at least four sources. In the next section, we describe the Census Tree and comparable datasets along three key dimensions: their size, their quality, and their representativeness.

IV. Results

A. Size (Recall)

In Figure 2, we compare the match rates for the CLP, MLP, Family Tree, and Census Tree for adjacent censuses. Starting with the rates for men in Figure 2A, we see that the Census Tree obtains allows us to create between 69% and 76% of the possible matches for the 19th century censuses, and between 82% and 86% for the 20th century. These exceptionally high match rates represent a large increase over existing linking methods. The Census Tree has five to six times as many links for men as the CLP (Exact-Conservative, or EC).[20] Comparing to the MLP, the Census Tree has between 41 and 80% more matches. Finally, Figure 2A shows the gain that is made by our process for expanding the Family Tree. The Census Tree dataset is 1.7 to 3 times larger than the Family Tree for these adjacent census pairs.[21]

[20] In Figure 2 we use the Exact-Conservative matches from the CLP. We choose this method when comparing match rates because it has precision estimates that are closest to those of the MLP and the Family Tree. Match rates are higher for other sets of the CLP links (reaching 30-40%).

[21] The Family Tree has the highest match rates for 1900-1910 and 1910-1920 because the Record Linking Lab at BYU has focused their initial efforts to expand the Family Tree on the 1910 census. The Lab's goal has been to ensure that every person in the 1910 census has a profile on the Family Tree, and as of July 2024, the coverage rate had reached 73%.

Match rates for women are in Figure 2B. The CLP has match rates of 0% for all years, as they do not attempt to link women. The MLP does, with rates between 32% and 46% for their adjacent-census pairs. The Census Tree's match rates are 1.6 to 1.9 times higher, and range from 58% to 79%, with all four 20th century pairs obtaining match rates above 70%. As with men, the Census Tree process adds millions of observations to those in the Family Tree, increasing the match rates by 50 to 300%. We note that the gain in going from the Family Tree to the Census Tree is slightly smaller for women than it is for men. This is because users link their female and male ancestors at very similar rates, but our XGBoost algorithm is not able to "learn" to make matches for women in cases where the surname changes due to marriage.

We include match rates for all 36 census-to-census pairs in Appendix Table 4. Here, we do not remove new immigrant arrivals from the linkable population because this adjustment performs poorly for censuses that are further apart in time.[22] Although these match rates are attenuated, the match rates for men are still above 56% for all census-to-census pairs. As expected, match rates are generally higher for more recent censuses. It is the case that the match rates are actually *above* 100% for pairs that are 80 or 90 years apart; this appears to be due to likely errors in the denominator (e.g. unreliable ages for those who are very old).[23] The match rates for women show similar patterns, with rates of 44% or above for all pairs, and again reaching 70% or above in the 20th century.

Table 4 translates these match rates into the number of links between each of the 36 census-to-census pairs. These numbers show the unprecedented size of the Census Tree dataset, with over 391 million links for men and 314 million links for women. While the calculation of the match rates is sensitive to choices about how the denominator is constructed, the absolute number of links is not. Accordingly, Table 4 also shows that the size of the crosswalks predictably declines as the length of time between the two censuses grows.

---

[22] As described in Price et al. (2021), our main match rate calculation adds the total number of reported immigrants in the U.S. between the two years, and subtracts this total from the denominator. When the censuses are further apart, this causes the denominator to be much too small, as many of those who immigrated between the two endpoints will not have survived to the latter year. Ideally we would use information on the year of immigration from the latter census to adjust the denominators, but this information is only available from 1900-1930.

[23] A large literature documents the presence of unreliable older ages in self-reported data, including in the US census (Rosenwaike 1979, Ewbank 1981, Preston et al. 1998). It is also the case that a higher fraction of links at longer intervals are identified only by the "implied" method. For example, 29% of the 1850-1940 links are made using this method alone, compared to 19% of 1900-1940 links, and 3% of 1930-1940 links. In the next section, we discuss ways that the researcher could use information about a link's source to choose a point on the precision-recall frontier.

B. Quality (Precision)

While it is clear that the Census Tree is an advance in terms of the number of links made, what can we say about whether the links are likely to be "true" matches? As we described in Section II.B., we randomly selected 1,000 people in the 1910 census for whom it should be possible to find a match to 1900. For the 759 with a Census Tree link to 1900, we asked research assistants to use the full set of information available in each census record to classify each link as correct, incorrect, or unsure. Table 5 shows the fraction of each links that were determined to be correct, for the full Census Tree and for the links identified by each link source. This fraction—known as precision—depends on the treatment of the "unsures," and so we present results with different treatments that constitute upper and lower bounds.

Between 89% and 94% of the links in the full Census Tree were determined to be correct, depending on whether the unsure links are considered incorrect, correct, or dropped altogether. When we look at the source of the links, we see that the implied links are the least precise and the Family Tree links are the most precise. The supervised methods (XGBoost, MLP, FS Hints) have very similar precision, and perform better than the unsupervised method (CLP). Note that each individual method has a higher rate of precision than the full Census Tree, because the precision for each method is calculated using two types of links—links that are only identified by that method, and links that are identified by that method *and* by other methods.[24]

In Table 5 we also compare precision for links that are identified by one or more sources. When a link is only identified by one source, it is determined to be correct between 69% and 82% of the time. However, links that have two sources are much more precise (86% to 94%). Links that have at least four sources have precision rates of 94% or above—reaching 99% and 100% for those with six and seven sources, respectively.

The results in Table 5 highlight the well-known tradeoff between recall and precision in record linking (Abramitzky, Boustan, and Eriksson et al. 2021). The Census Tree constitutes a major advance in what is possible in terms of match rates, while maintaining high rates of precision. However, for some applications, researchers may prefer to have even higher confidence in the

[24] This is possible if links that are identified by two or more methods are more precise than other links. To see this, suppose that there are two methods (X and Y) and six links. Two of the links are identified by X only, one of which is correct. Two are identified by Y only, and again one is correct. Finally, two links are identified by both X and Y, and both are correct. While the precision for each method would be 0.75 (3/4), precision for the entire set would be 0.67 (4/6) because the more precise links identified by both X and Y comprise a smaller fraction of the set.

matches even if it means reducing their sample size. For this reason, the Census Tree crosswalks include flags that indicate the sources of the match. With these flags, the researcher could omit links from methods that they believe to be lower quality (e.g. the implied links at longer intervals). A researcher could also use the flags to require that the links be identified by a minimum number of sources. In Figure 3, we illustrate the tradeoff that one would be making when imposing such a restriction, using the 1900-1910 links. When all 47.4 million links between these two censuses are included, we estimate that 93.9% are correct.[25] Restricting the sample to links that are identified by *at least* two sources would increase precision to 96.6%, but would reduce the sample size by about 17% (still leaving 39.4 million links). Further requiring that at least three sources identify the link will increase precision to 97.1% and will leave 33.0 million observations. From there, the tradeoff becomes roughly linear. One-hundred percent of links identified by all seven sources are expected to be correct, but the sample size would be just 3.1 million. Conceptually, the points in Figure 3 trace out a "production possibilities frontier," and the source flags in the publicly available Census Tree crosswalks allow the researcher to choose their desired point along this frontier.

C. Representativeness

Another desirable property of any dataset is that it be representative of the population it is meant to describe. This has been a challenge for those attempting to create linked datasets, as some people may be easier to link, leading to selected samples (Bailey et al. 2020). The most serious issue has been the difficulty in linking women, but other populations that have been difficult to link include those with common names, those whose names are less stable (e.g. immigrants), or those who are more likely to have been left out by the enumeration process (e.g. the enslaved or formerly enslaved) (Hacker 2013).

To assess the representativeness of the Census Tree and its alternatives, we compare the observable characteristics of those linked between 1900 and 1910 by each method to the full population of those who are observed in the 1910 census. From the latter, we omit those who are under age 11, as those children would not have been born in 1900. The results are in Table 6 (see Appendix Table 5 for comparisons between the Census Tree and the population for other adjacent

[25] Precision estimates are from the quality check used to create Table 5. In Figure 3 we use the precision estimates where the "unsures" are dropped, but the tradeoff is very similar if those links are treated as either correct or incorrect. Note that in Table 5, we estimate precision when a link is identified by *exactly* N sources; in Figure 3, we estimate precision when a link is identified by *at least* N sources, which is consistent with the exercise described here.

year pairs). As expected, the Census Tree has nearly the same fraction of women as the population (0.47 vs. 0.48), compared to 0.43 for the MLP and zero for the CLP. As with previous efforts, the fraction Black is lower than that of the full population, but our process yields a significant improvement on this front relative to the Family Tree alone. Furthermore, the large sample size in the Census Tree means that there are still 3.39 million links for Black Americans between 1900 and 1910—over 1.5 million more than are available in the MLP.

Those linked in the Census Tree are very similar to the full population in terms of their marital status and family structure. There is some evidence that those on the Census Tree are positively selected by socioeconomic status—they are slightly more literate and more likely to speak English. They are also more likely to live in their birth state. On all of these dimensions, the Census Tree does at least as well at matching the population as the CLP, the MLP, or the Family Tree alone.

Critically, the summary statistics in Table 6 and Appendix Table 5 are unweighted. Bailey, Cole, and Massey (2020) propose a method for weighting linked data to match population characteristics and obtain representative samples. Buckles et al. (2023) apply their method and show that, once weighted, estimates of the intergenerational transmission of socioeconomic status are nearly identical when using links from either the CLP or the Census Tree, despite the fact the two datasets have different sample sizes and observable characteristics. Moreover, the Census Tree has such large samples that the reweighting procedure is likely to have sufficient support in the data for reweighting in cases where the study population is smaller (e.g. a single state or immigrant group).

To summarize, there is little evidence that the Census Tree dataset is a highly selected sample—as we would expect, given that each year-to-year pair has at least 60% of the linkable population. Where some non-representativeness remains, the dataset is large and complete enough to support re-weighting to produce results that match the population characteristics. The Census Tree also includes millions of observations for groups that have been omitted or under-represented in prior research, including women and the formerly enslaved and their descendants.

V. Application: Intergenerational Mobility of Immigrants and the U.S.-born

With its combination of a high match rate, high precision, and a representative sample that includes women, the Census Tree will allow researchers to answer new questions and to improve and extend prior work. As an example of the latter, we revisit an influential paper by Abramitzky, Boustan, Jácome, and Pérez (2021)—or ABJP—on the intergenerational mobility of immigrants. In this paper, the authors compare historical income mobility estimates (using an occupation-based predicted income) for White sons of immigrant fathers and sons of U.S.-born White fathers. The key finding is that sons of immigrant fathers have higher rates of upward mobility on average than sons of the native born. The authors provide mobility estimates for each of the 17 largest sending countries and conclude that, for nearly all of them, the son of an immigrant father at the 25th percentile of the earnings distribution is expected to attain a higher income rank than the son of a similar U.S.-born White father.

In this section, we begin by replicating several key findings in ABJP using the Census Linking Project (CLP) data. We then repeat the exercise using the Census Tree, using the same sample restrictions. Estimates of intergenerational mobility are nearly identical between the Census Tree and the CLP, but the Census Tree estimates are more precise because the they are produced using a considerably larger sample. Next, we extend the analysis presented in ABJP in two ways. First, we leverage the Census Tree's higher match rate to produce estimates of intergeneration mobility for several more sending countries and for sons of Black U.S.-born fathers. Second, because the Census Tree includes links for women as well as for men, we produce novel estimates for daughters of immigrant and U.S.-born fathers.

A. Reproduction of ABJP Using the Census Linking Project

We begin by reproducing the main results in ABJP using the code from their replication package (Abramitzky et al. 2021b) and the links from the CLP. We focus on results for the 1880-1910 cohort and include results from the 1910-1940 cohort in the appendix. Following ABJP, we restrict the sample to White sons ages 0-16 in 1880 who are living with their father ages 30-50. The immigrant sample is restricted to fathers from the 17 largest sending countries in 1880. In Figure 4 Panel A, we plot rank-rank estimates of intergenerational mobility across percentiles of the father's income distribution, following Figure 2 in ABJP. There are small differences in the slope and intercept estimates in the two figures because the ABJP replication code uses restricted-use census data to create the links directly, while we use the publicly available CLP links. But critically, we reproduce the key finding—the sons of immigrants do better than the sons of White U.S.-born fathers on average, with larger gaps at lower income levels.

In Figure 5 Panel A, we show the average predicted rank in the income distribution for sons with fathers in the 25th percentile, for the 17 largest sending countries and the U.S. (following ABJP's Figure 3). Here again, we replicate the main conclusion—the sons of U.S.-born fathers attain a lower point in the income distribution than sons of immigrants from nearly every country (except

Norway and Belgium). There are small differences in the ordering of the countries, but we also find the highest mobility among sons of fathers from Ireland and Portugal.

B. Replication of ABJP Using the Census Tree

Having confirmed that we are able to reproduce the main findings of ABJP using the CLP, we now repeat this analysis using the Census Tree links and compare the estimates. Beginning with Figure 4 we see that the intercept and slope estimates for both White U.S.-born fathers and immigrant fathers are very similar when using the CLP links (Panel A) and the Census Tree links (Panel B). We also include new estimates of the intergenerational mobility of sons of Black U.S.-born fathers in Panel B, using the large set of Census Tree links for Black Americans. The regression intercept for this group is dramatically lower (at 16.73) than for sons of White U.S.-born or immigrant fathers. The regression slope is also much flatter, at 0.12.

In Figure 5 Panel B, we replicate Panel A using the Census Tree. We find that the sons' earning percentile rank is similarly ordered across countries in the two panels. In the CLP sample, the sons of low-income U.S.-born fathers attain a lower point in the income distribution than similar sons of immigrants from every country except Norway and Belgium. In the Census Tree sample, the estimate for Belgium is slightly above that for the U.S., though the point estimates are not statistically distinguishable in either sample. Sons from Ireland and Portugal have the highest upward mobility in both samples. There is some re-ordering of the countries in the middle range of the mobility estimates (Switzerland, Austria, Germany, and Finland), but the point estimates for these countries are tightly clustered in both panels. Results using the Census Tree are also similar when compared directly to Figure 3 in ABJP.

Each estimate from the Census Tree is produced using a larger sample of father-son pairs, which yields more precise point estimates. To demonstrate this, we add 95% confidence intervals to the country-level estimates in Figure 5. Comparing Panel A and Panel B, the Census Tree sample has confidence intervals that are 24% smaller on average.

C. Extension: Additional Sending Countries

In ABJP, the country with the smallest number of links was Finland, with 90 observations in 1880. If we use this same threshold in the Census Tree, we can produce mobility estimates for an additional ten countries—Australia, the Bahamas, Cuba, Czechoslovakia, Hungary, Luxembourg, Mexico, the Netherlands, Poland, and Spain. Crucially, three of these countries are from Central America and the Caribbean, which allows us to examine the mobility of immigrants from this part of the world. These new estimates are included in Figure 5, Panel C. A key finding is that the sons of White immigrants at the 25th percentile from Mexico experience much lower mobility than the sons of similar immigrants from other countries and the White U.S.-born. This is also true for the 1910 cohort (Appendix Figure 3, Panel C). Returning to the 1880 cohort, the sons of low-income immigrants from Cuba and Luxembourg are also expected to attain lower income ranks than the sons of the White U.S.-born. Meanwhile, the newly-added Hungary and Poland rank among the five most mobile sending countries.

While these new findings support ABJP's main conclusion that the children of immigrants at the 25th percentile are more mobile than the children of the White U.S.-born, they add important nuance to this result. In particular, we show that the U.S. was not a "Land of Opportunity" for the children of low-income Mexican immigrants in the late 19th and early 20th centuries in the same way that it was for other groups. This contrasts with the findings in ABJP using more recent data from Opportunity Insights (Chetty et al. 2018), in which estimates for the 1997-2015 cohorts show that expected incomes for children of low-income fathers are higher for Mexico than for the U.S. Thus, the relatively high mobility of the children of low-income Mexican immigrants appears to be a recent phenomenon. This is consistent with Kosack and Ward (2020) who document a mobility gap for Mexican Americans relative to similar peers in the U.S. in the late 19th century that narrows by the end of the 20th century.

We also extend the analysis in ABJP by adding estimates for the sons of Black U.S.-born fathers. ABJP omitted estimates for sons of Black U.S.-born fathers to ensure "that the higher mobility of second-generation immigrants that we observe is not due to Black-White differences in mobility" (ABJP, p. 581). We include them here to enable us to compare this group to those from Central America and the Caribbean, which also had large non-White populations. We find that the sons of low-income U.S.-born Black fathers had much lower mobility than the sons of similar White fathers—a result that is consistent with Collins and Wanamaker (2022). Moreover, Black sons in this sample also had worse outcomes on average than the sons of White immigrants from any of the top 27 sending countries, including Mexico. The gap between sons of U.S.-born Black men and immigrants from Mexico widened for the 1910 cohorts (Appendix Figure 3). For both cohorts, the sons of U.S.-born Black men at the 25th percentile of the income distribution are actually predicted to have incomes that are *below* the 25th percentile.

D. Extension: Estimates for Women

A signature contribution of the Census Tree dataset is the fact that it includes women, including millions of links between childhood and adulthood. This allows us to further extend ABJP by producing estimates for daughters; ABJP do this for the more recent 1997-2015 cohorts but not for historical samples. We follow prior work and use a married woman's husband's status as a measure of her own (Olivetti and Paserman, 2015); we discuss the implications of this choice below.

In Panel C of Figure 4, we show that the rank-rank correlations for married daughters are very similar to those of men, a finding consistent with Buckles et al. (2023) and with other recent work documenting a high degree of assortative mating in this period (Curtis 2021; Clark and Cummins 2022; Olivetti et al. 2020). We can compare these new historical estimates of rank-rank correlations for daughters from the 1880-1910 and 1910-1940 cohorts to those for the 1997-2015 Opportunity Insights cohorts in ABJP. In the latter, the daughters of immigrant men attain higher ranks in the income distribution on average than the daughters of U.S.-born men across the income distribution, though the slope is identical for the two groups (0.25). In the historical samples, the slope is steeper for both groups—as was the case for sons in ABJP. Furthermore, the slope for the daughters of the U.S.-born is steeper than that for the daughters of immigrants (0.34 vs. 0.28 for the 1880-1910 cohorts), so that at the higher end of the distribution there is almost no gap in expected income rank between the two groups. An important caveat is that the estimates for the historical samples use the son-in-law's income in place of the daughter's, while the estimates from the Opportunity Insights cohorts use the daughter's own income.

In Panel D of Figure 5, we show the average daughter's (son-in-law's) place in the income distribution for those with fathers at the 25th percentile, for the same set of countries for men in Panel C. Again, the main conclusion holds about the relative position of daughters of U.S.-born White and U.S.-born Black men compared to daughters of immigrants, though there are changes in the ordering. The most significant change occurs for children of low-income fathers who immigrated from Cuba: their sons attain an income rank around the 40th percentile on average, while their daughters marry men at the 55th percentile. This example highlights the role that marriage patterns play in generating estimates for daughters when husbands' earnings are used to measure their socioeconomic status. First, the children of Cuban immigrants had relatively low marriage rates—only 68% of Cuban daughters and 71% of Cuban sons in these cohorts were married in 1910. If there is positive selection into marriage, we would expect that the married daughters of Cuban immigrants would appear to be more upwardly mobile when only married women can be included in the sample. Second, the daughters of Cuban immigrants were more likely to marry the sons of immigrants from other countries than were daughters from other immigrant groups (perhaps owing to their relatively small number in this time period). For example, 10% of the daughters of Cuban immigrants married sons of men from Spain, which is one of the more upwardly mobile countries for men in Figure 5, Panel C. Thus, intermarriage may have been a way for the daughters of low-income fathers from Cuba to improve their station. We leave a more in-depth investigation into the historical relationship between marriage patterns and intergenerational mobility for future work—work made possible by the Census Tree.

VI. Conclusion

The Census Tree—which is available for download at censustree.org—is a resource that allows researchers to link people across the historical United States censuses at an unprecedented scale. Scholars will be able to create longitudinal datasets that follow individuals over time, and to connect people to their families and communities. In this paper, we have described our process for creating this resource, which incorporates links provided by the users of an online genealogy platform. We then add additional links using machine learning and the contributions of previous linking efforts, and adjudicate conflicts among various link sources. The finished dataset contains over 700 million links, including 314 million links for women and 41 million links for Black Americans. The Census Tree flexibly accommodates different preferences regarding the tradeoff between recall and precision, and it is large enough to support reweighting and research on small populations.

Our hope is that the Census Tree will allow researchers to answer new questions in history and the social sciences. In fact, this is already happening—in Appendix B, we include a bibliography of published and working papers that have used the Census Tree links. Given that a key innovation is the inclusion of women in large numbers, it is unsurprising that several of these papers focus on topics related to women and gender. For example, Aneja et al. (2024) study the intergenerational effects of exposure to women employed during World War I on gender norms; Bazzi et al. (2023) link modern-day gender norms to the frontier experience; Abramitzky et al. (2023) examine the gendered impacts of skin tone; Jones et al. (2023) document parents' gender preferences from 1850 to the present, and Vidart (2024) revisits the link between electrification and fertility. Others use the Census Tree to create long panels that allow them to study the long-term impacts of policies and environments, such as exposure to regional universities (Howard and Weinstein 2024), agricultural education (Minyo 2024), charity nurseries (Ager and Melin 2024), family structure (Cools et al. 2024), and parental socioeconomic status (Buckles et al. 2023; Ellsworth et al. 2024). Gabriel et al. (2023) show that family members of lynching victims were more likely to migrate during the period of the Great Migration. [26]

The Census Tree will also allow researchers to replicate and extend prior work. To demonstrate this, we explore the findings of Abramitzky, Boustan, Jácome, and Pérez (2021), who show that the children of immigrants have historically had higher socioeconomic mobility on average than similar children of the U.S.-born. We are able to replicate this result using the Census Tree, and we increase the precision of the estimates in the original paper. Furthermore, we add estimates for ten new immigrant-sending countries, including from Central America and the Caribbean. While the main conclusions of ABJP hold, we show that the sons of immigrants from Cuba and especially Mexico did not experience higher mobility than the sons of the U.S.-born for our cohorts. Furthermore, we are able to produce novel historical estimates of socioeconomic mobility for married women, using their husband's income rank as a proxy for their own. We find that overall mobility estimates for sons and daughters are very similar, but that marriage patterns contribute to gender differences in estimates for some sending countries. Thus, this paper enriches our understanding of the relative mobility of immigrants and the U.S.-born in the 19th and 20th centuries.

[26] See Appendix B for bibliography of papers cited in this paragraph.

References

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

Abramitzky, Ran, Leah Boustan, Elisa Jácome, and Santiago Pérez. 2021. "Intergenerational Mobility of Immigrants in the United States over Two Centuries." *American Economic Review*, 111 (2): 580-608. https://doi.org/10.1257/aer.20191586

Abramitzky, Ran, Leah Boustan, Elisa Jácome, and Santiago Pérez. 2021b. Data and Code for: Intergenerational Mobility of Immigrants in the US over Two Centuries. Nashville, TN: American Economic Association [publisher]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-01-28. https://doi.org/10.3886/E120490V1

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez, and Myera Rashid. 2020. "Census Linking Project: Version 2.0 [Dataset]. https://censuslinkingproject.org.

Bailey, Martha, Connor Cole, and Catherine Massey. 2020. "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850–1930 IPUMS Linked Representative Historical Samples." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (2): 80–93. https://doi.org/10.1080/01615440.2019.1630343.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature* 58 (4): 997–1044. https://doi.org/10.1257/jel.20191526.

Berkes, Enrico, Ezra Karger, and Peter Nencka. 2023. The Census Place Project: A Method for Geolocating Unstructured Place Names. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E179401V2.

Black, Sandra E, Paul J Devereux, and Kjell G Salvanes. 2005. "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital." *American Economic Review* 95 (1): 437–49. https://doi.org/10.1257/0002828053828635.

Buckles, Kasey, Joseph Price, Zach Ward, and Haley Wilbert. 2023. "Family Trees and Falling Apples: Intergenerational Mobility Estimates from U.S. Genealogy Data." Working paper. https://www.kaseybuckles.com/s/BHPW_aug2023.pdf.

Chetty, Raj, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. *"Race and Economic Opportunity in the United States: An Intergenerational Perspective."* NBER Working Paper #w24441.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106 (4): 855–902. https://doi.org/10.1257/aer.20150572.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States*." *The Quarterly Journal of Economics* 129 (4): 1553–1623. https://doi.org/10.1093/qje/qju022.

Collins, William J., and Marianne H. Wanamaker. 2022. "African American Intergenerational Economic Mobility since 1880." *American Economic Journal: Applied Economics* 14 (3): 84–117. https://doi.org/10.1257/app.20170656.

Ewbank, Douglas C., 1981. *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis* (Vol. 4). National Academy Press.

FamilySearch. 2023. "The Largest Free Family History Resource." FamilySearch. 2023. https://www.familysearch.org/en/about/.

———. 2023. "Who Are Your Ancestors?" 2023. https://ancestors.familysearch.org/en/.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking," 34.

Feigenbaum, James J. 2018. "Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940." *The Economic Journal* 128 (612): F446-F481.

Fogel, Robert W., and Larry T. Wimmer. 1992. "Early indicators of later work levels, disease, and death." NBER Working Paper #h0038.

Folkman, Tyler, Rey Furner, and Drew Pearson. 2018. "GenERes: A Genealogical Entity Resolution System." In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 495–501. https://doi.org/10.1109/ICDMW.2018.00079.

Gross, Matthew, and Michael Mueller-Smith. 2021. "Modernizing Person-Level Entity Resolution with Biometrically Linked Records." https://matthew-gross.github.io/entity_resolution_20210628.pdf

Hacker, J. David. 2013. "New Estimates of Census Coverage in the United States, 1850-1930." *Social Science History* 37 (1): 71-101.

Helgertz, Jonas, Steven Ruggles, John Robert Warren, Catherine A. Fitch, David Hacker, Matt Nelson, Joseph Price, Evan Roberts, and Matthew Sobek. 2023. "IPUMS Multigenerational Longitudianal Panel: Version 1.1 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D016.V1.1.

Kaplanis, Joanna, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, et al. 2018. "Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives." *Science* 360 (6385): 171–75. https://doi.org/10.1126/science.aam9309.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics* 11 (4): 181–209. https://doi.org/10.1257/app.20180010.

Kosack, Edward. and Ward, Zachary. 2020. "El Sueño Americano? The Generational Progress of Mexican Americans Prior to World War II." *The Journal of Economic History*, *80*(4): 961-995.

Mazumder, Bhashkar. 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *The Review of Economics and Statistics* 87 (2): 235–55. https://doi.org/10.1162/0034653053970249.

Olivetti, Claudia, and M. Daniele Paserman. 2015. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." *American Economic Review*, 105 (8): 2695–2724.

Olivetti, C., Paserman, M.D., Salisbury, L. and Weber, E.A. 2020. *Who married,(to) whom, and where? Trends in marriage in the United States, 1850-1940*. NBER Working Paper #28033.

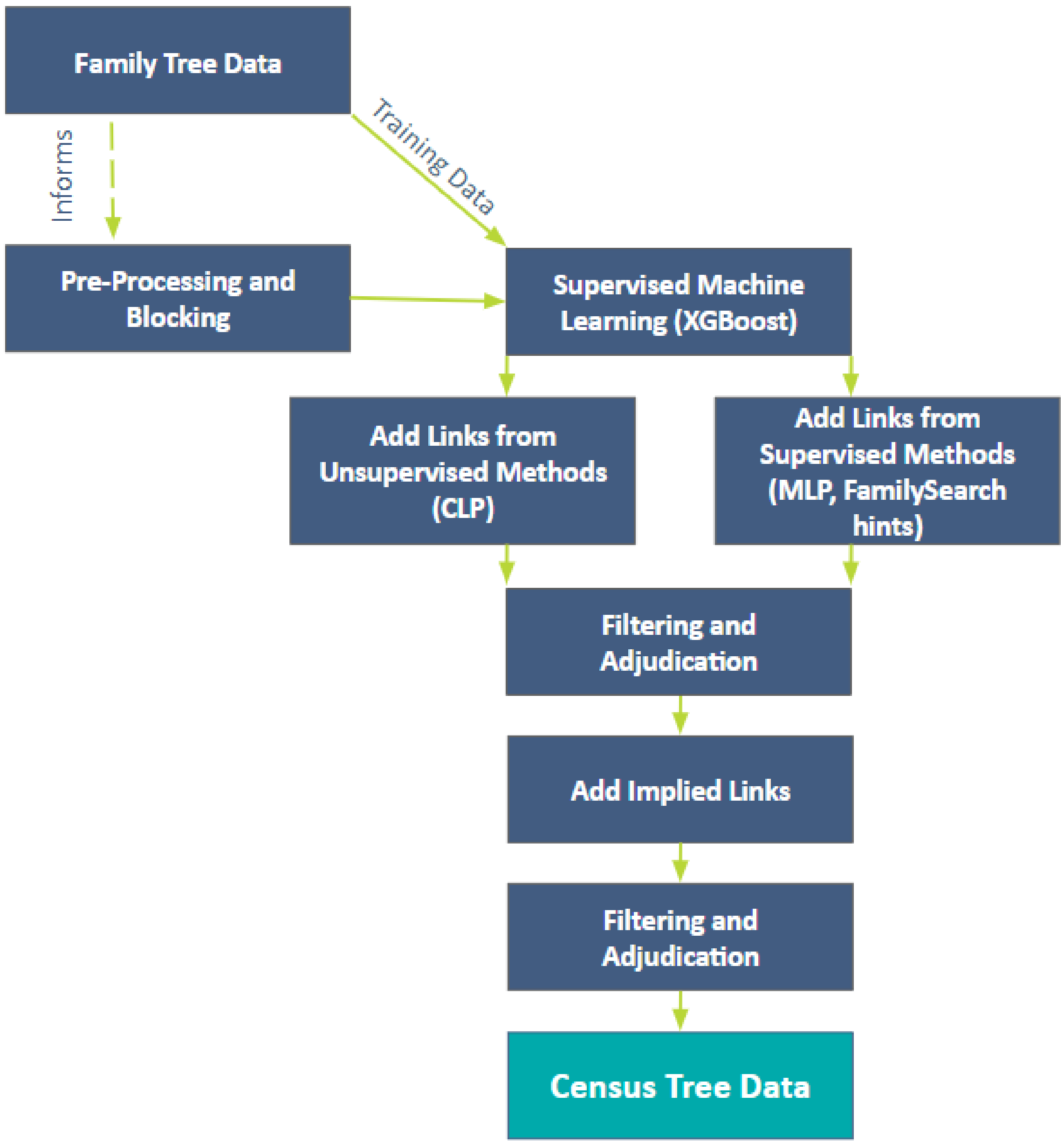
Preston, Samuel H., Elo, Irma T., Foster, Andrew and Fu, Haishan. 1998. "Reconstructing the Size of the African American Population by Age and Sex, 1930–1990." *Demography*, *35*(1), pp.1-21.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set." *Explorations in Economic History* 80 (April): 101391. https://doi.org/10.1016/j.eeh.2021.101391.

Rosenwaike, Ira. 1979. "A New Evaluation of United States Census Data on the Extreme Aged." *Demography*, *16*(2), pp.279-288.

Ruggles, Steven, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Evan Roberts, Megan Schouweiler, and Matthew Sobek. 2021. IPUMS Ancestry Full County Data: Version 3.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D014.V3.0

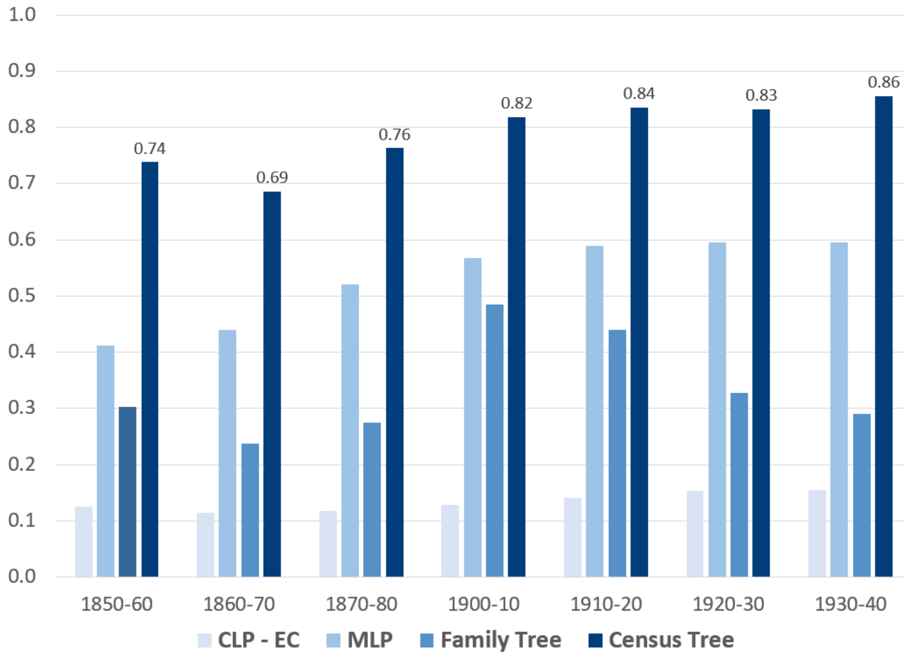
Figure 1: The Process for Creating the Census Tree



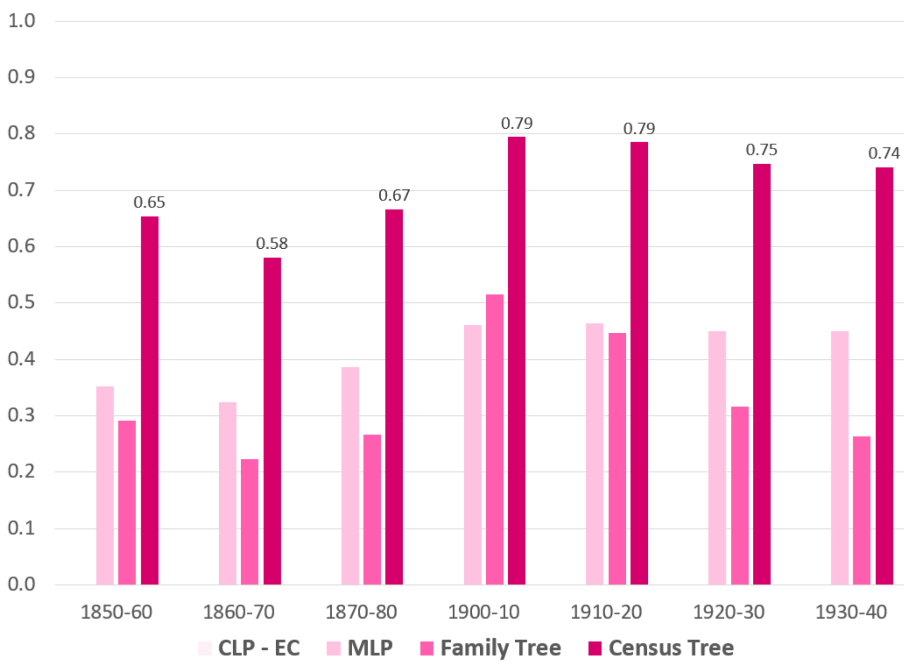
Notes: CLP links are from the Census Linking Project, using the NYSIIS standard links. MLP links are from the IPUMS Multigenerational Longitudinal Panel, and FamilySearch hints are created by FamilySearch using their proprietary algorithm. See the text for a description of implied links and of the filtering and adjudication process.

Figure 2: Match Rates Using Various Linking Methods, for Censuses Ten Years Apart

Panel A: Men

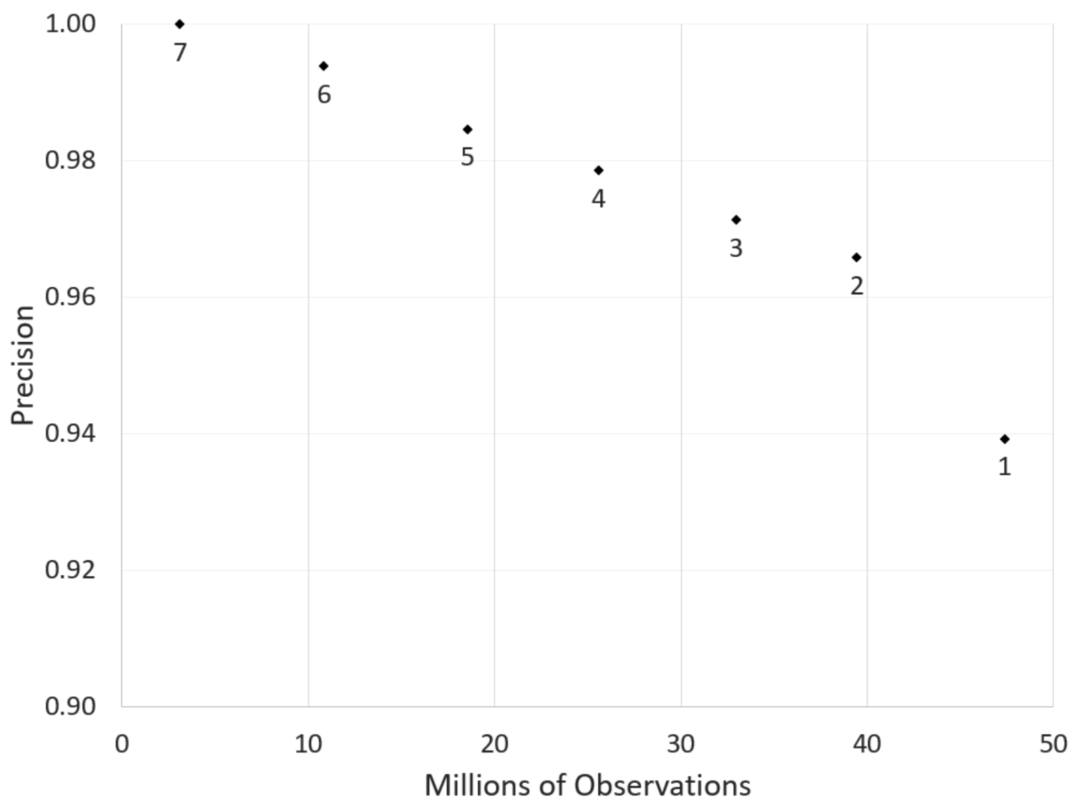


Panel B: Women



Notes: Match rates are constructed as the number of links between the two years, divided by the number of people age 11 and older in the latter year, with adjustment for rates of under-enumeration in the earlier census and for immigration. CLP – EC links are from the Census Linking Project, using the exact conservative approach; the CLP match rate for women is 0% for all pairs, as the CLP does not attempt to link women. MLP links are from the IPUMS Multigenerational Longitudinal Panel. Family Tree links are made by users on FamilySearch.org, and the Census Tree links are from the final Census Tree dataset.

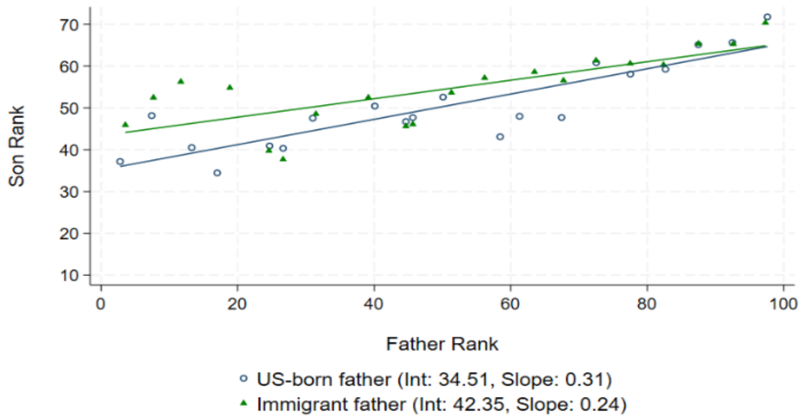
Figure 3: Precision and Sample Size, Conditional on Number of Link Sources, 1900-1910



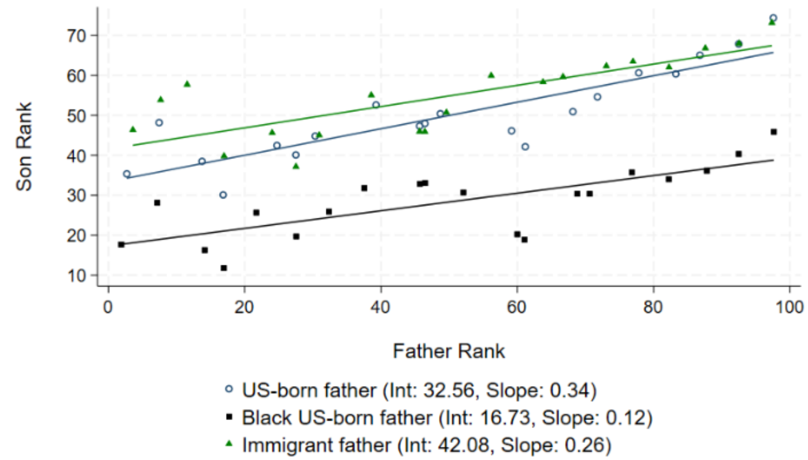
Notes: The figure shows the number of links and estimated precision (fraction of links expected to be correct) for links from 1900-1910, when the sample is limited to links identified by at least *n* sources ($1 \leq n \leq 7$). Precision estimates are from the quality check used to create Table 5, where links classified as "unsure" are dropped.

Figure 4: Rank-Rank Estimates of Intergenerational Mobility for 1880-1910, Replicating ABJP (2021)

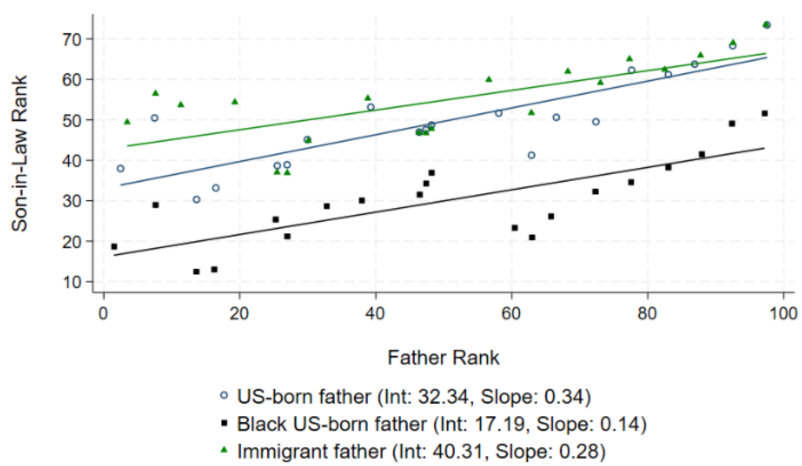
Panel A: Replication Using CLP (Men)



Panel B: Using Census Tree (Men)



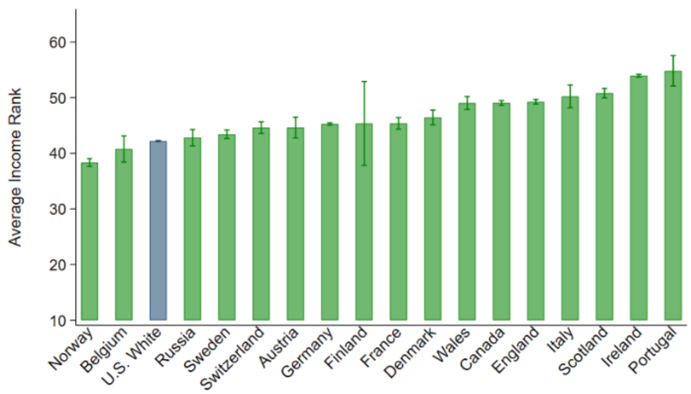
Panel C: Using Census Tree (Women)



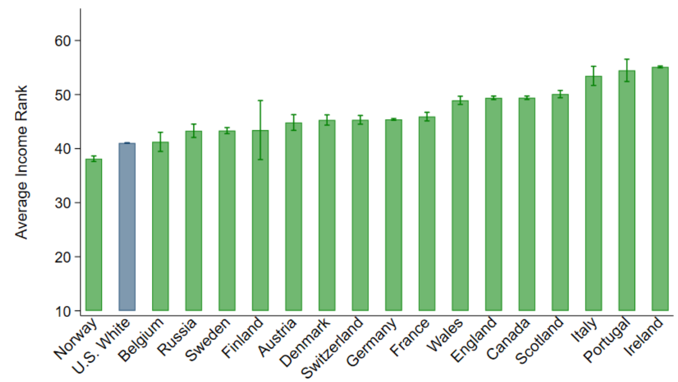
Notes: Panel A replicates the rank-rank estimates of intergenerational mobility for the 1880-1910 cohort in Figure 2 of Abramitzky, Boustan, Jácome, and Pérez (2021). Data are from the Census Linking Project. In Panel B, we produce the same estimates using the Census Tree, and add estimates for Black men. In Panel C, we use the Census Tree to produce estimates for women.

Figure 5: Average Income Rank for Children Born to 25th Percentile by Father's Birthplace, 1880-1910 Cohort, Replicating ABJP (2021)

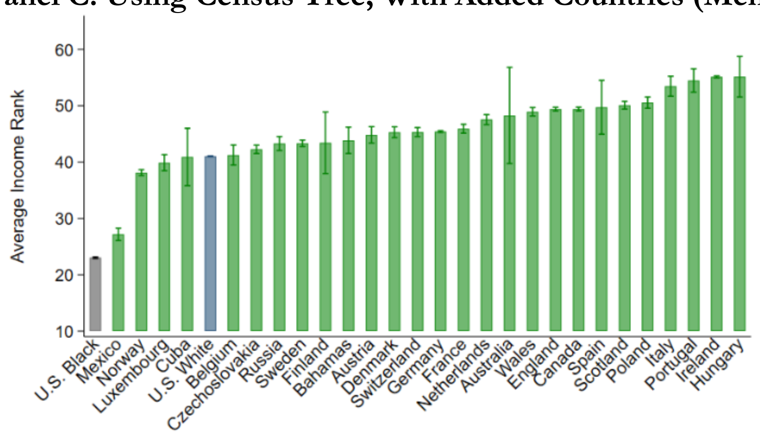
Panel A: Replication Using CLP (Men)



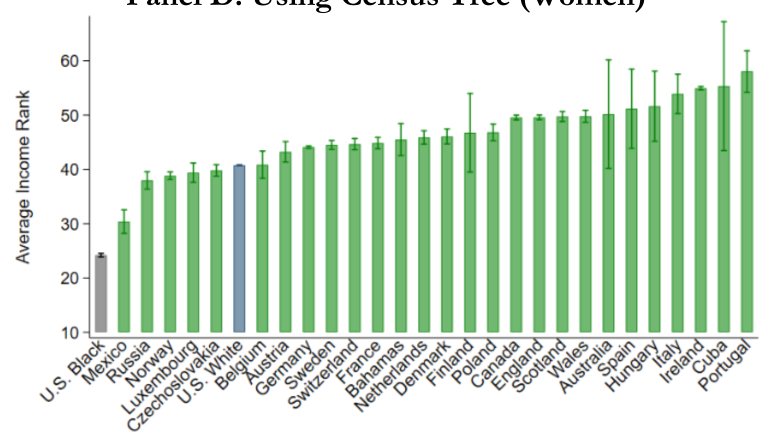
Panel B: Using Census Tree, ABJP Countries (Men)



Panel C: Using Census Tree, With Added Countries (Men)



Panel D: Using Census Tree (Women)



Notes: Panel A replicates the estimates of intergenerational mobility at the 25th percentile for the 1880-1910 cohort in Figure 3 of Abramitzky, Boustan, Jácome, and Pérez (2021). Data are from the Census Linking Project, and estimates are shown for U.S.-born Whites and for the 17 largest sending countries. In Panel B, we produce the same estimates using the Census Tree. In Panel C, we add 10 countries with samples large enough to be identified in the Census Tree. We also add estimates for Black men. In Panel D, we replicate Panel B for women. The figure shows 95% confidence intervals for each estimate.

Table 1: Fifteen Most Important Features Used by XGBoost Algorithm, 1900-1910

Feature	Description	Importance
Township distance	Geographic distance between townships	0.1890
Birth year difference	Absolute difference in birth years	0.1047
Middle initial exact	Indicator for middle name exact match	0.0961
Last name uniqueness * last name Levenshtein	Levenshtein string distance in last name, weighted higher for more unique names	0.0722
Last name uniqueness * last name exact	Indicator for last name exact match, weighted higher for more unique names	0.0606
Sign of birth year difference	Sign of difference in birth years	0.0452
Mother's birthplace exact	Indicator for mother's birthplace exact match	0.0383
First name uniqueness * first name Jaro-Winkler	Jaro-Winkler string distance in first name, weighted higher for more unique names	0.0367
State exact * not living in birth state	Indicator for residence state exact match and living outside birth state	0.0304
Immigrant in starting year	Indicator for immigrant in 1900	0.0277
Standardized first name uniqueness * Standardized first name Levenshtein	Levenshtein string distance in standardized first name, weighted higher for more unique names	0.0264
Last name Jaro-Winkler	Jaro-Winkler string distance in last name	0.0246
Relationship exact	Indicator for relationship to head exact match	0.0220
First name uniqueness * first name Levenshtein	Levenshtein string distance in first name, weighted higher for more unique names	0.0214
Father's birthplace exact	Indicator for father's birthplace exact match	0.0212

Notes: The importance measure is calculated as the average increase in accuracy across nodes of the decision tree which use the feature. This is the "gain" measure of feature importance calculated by the XGBoost algorithm. The model used 70 features in total.

Table 2: Number of Links in Census Tree from Each Source, 1900-1910

	Links Before F&A	% Dropped in F&A	In Census Tree	Unique Links
Family Tree	29,314,798	1.5%	28,874,030	672,841
XGBoost	27,407,692	7.6%	25,317,190	1,470,857
CLP	10,140,318	17.3%	8,388,152	406,770
MLP	30,313,883	1.9%	29,730,141	2,069,840
FS Direct Hint	26,963,154	1.6%	26,534,259	485,118
FS Profile Hint	26,455,508	3.3%	25,589,016	502,274
Implied Links	35,461,926	5.6%	33,468,423	2,314,368

Notes: F&A refers to the filtering and adjudication process described in the text. CLP refers to the Census Linking Project, MLP is the Multigenerational Longitudinal Panel from IPUMS, and FS is FamilySearch. See the text for descriptions of these linking methods.

Table 3: Number of Sources that Identify Each Link, 1900-1910

# Sources	Links
1	7,922,068
2	6,486,142
3	7,369,745
4	7,039,613
5	7,698,195
6	7,727,108
7	3,126,507
Total	47,369,378

Table 4: Number of Links Between Each Census Pair in the Census Tree

Panel A: Men

	1850	1860	1870	1880	1900	1910	1920	1930
1860	5,953,069							
1870	4,807,145	8,039,672						
1880	4,333,197	7,123,099	11,782,599					
1900	2,695,500	4,929,583	8,200,784	13,335,316				
1910	1,747,473	3,677,947	6,594,217	11,171,366	24,976,021			
1920	871,800	2,369,744	4,845,756	8,897,192	20,404,710	30,164,025		
1930	260,564	1,207,167	3,179,682	6,656,256	17,397,956	25,541,131	35,888,058	
1940	35,159	366,183	1,607,101	4,318,697	13,941,716	21,358,133	29,849,380	42,665,479

Panel B: Women

	1850	1860	1870	1880	1900	1910	1920	1930
1860	4,957,966							
1870	3,645,218	6,731,066						
1880	3,181,946	5,497,251	9,946,288					
1900	1,826,668	3,424,974	5,897,591	10,712,054				
1910	1,159,058	2,426,824	4,469,124	8,329,400	22,384,230			
1920	600,132	1,533,262	3,149,807	6,333,399	16,832,433	26,957,112		
1930	199,283	791,564	1,992,939	4,533,001	13,312,424	20,314,679	31,584,817	
1940	31,704	276,303	1,061,350	2,988,184	10,411,260	15,925,055	23,141,851	37,516,261

Notes: Table shows the number of links between each of the 36 census-to-census pairs. There are 391,205,308 links for men and 314,083,062 links for women, for 705,288,370 total links. See Appendix Table 4 for match rates.

Table 5: Precision Estimates for the 1900-1910 Census Tree

	Treat Unsure as Incorrect (N = 759)	Drop Unsure (N = 718)	Treat Unsure as Correct (N = 759)
Record Source:			
CLP	0.890	0.958	0.961
MLP	0.942	0.967	0.968
XGBoost	0.924	0.973	0.975
Family Tree	0.975	0.977	0.977
FS Direct Hint	0.959	0.974	0.975
FS Profile Hint	0.962	0.969	0.970
Implied Link	0.904	0.941	0.944
Number of Sources:			
1	0.691	0.794	0.821
2	0.857	0.938	0.943
3	0.902	0.948	0.951
4	0.947	0.964	0.965
5	0.949	0.969	0.970
6	0.991	0.991	0.991
7	1.000	1.000	1.000
Full Census Tree	0.894	0.939	0.942

Notes: Table shows the results of an exercise in which research assistants hand-checked a sample of 1900-1910 links from the full Census Tree. To construct the sample, we randomly selected 1,000 people in the 1910 census for whom it should be possible to find a match to 1900. For the 759 with a link, the research assistants classified each as correct, incorrect, or unsure. The top panel shows results by record source, where a record can have multiple sources. The bottom panel shows the results by the number of sources that identified the link. In the first column the unsure links are treated as incorrect, in the middle they are dropped, and in the last they are treated as correct.

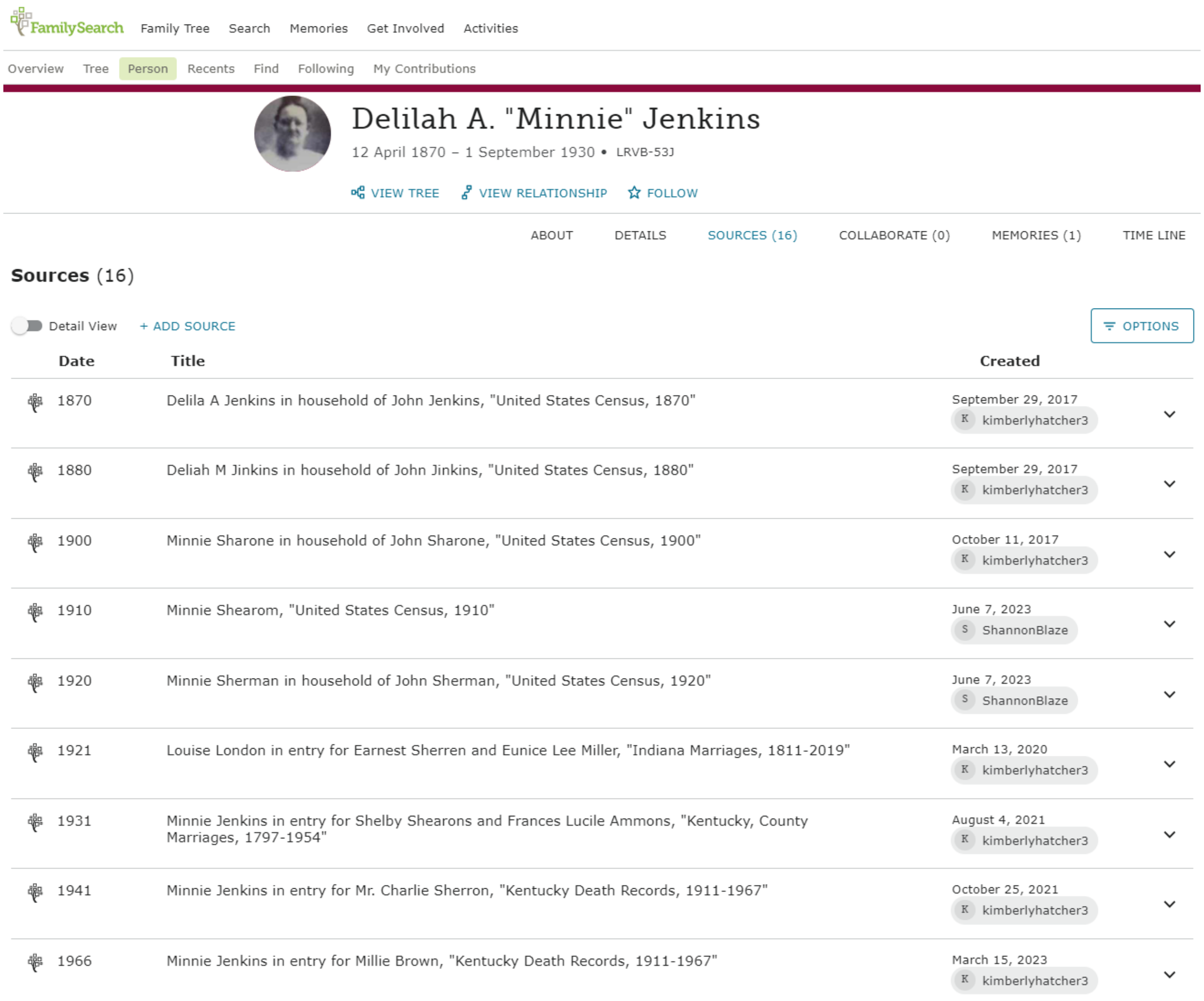
Table 6: Representativeness for Various Linking Methods, 1900-1910

	CLP	MLP	Family Tree	Census Tree	Full Census (Age 11+)
Female	-	0.4273	0.4947	0.4714	0.4824
Age	33.58	33.62	33.02	34.22	33.59
White	0.9239	0.9377	0.9451	0.9248	0.8925
Black	0.0742	0.0619	0.0544	0.0740	0.1030
Married	0.4912	0.4874	0.5317	0.5198	0.5133
HH Head	0.4901	0.2928	0.2844	0.3069	0.2876
HH Size	5.71	6.05	5.93	5.72	5.79
Lives in Birth State	0.6650	0.6934	0.7062	0.6671	0.5905
Speaks English	0.9859	0.9860	0.9901	0.9844	0.9501
Literate	0.9463	0.9501	0.9529	0.9425	0.9150
N	9,806,617	29,238,890	28,267,717	45,772,617	69,725,595

Notes: Unweighted summary statistics for individuals observed in 1910, for which each data set has a link for 1900, compared to the population of individuals age 11 or older in 1910. CLP links are the NYSIIS-Standard links from the Census Linking Project; the CLP does not include women. MLP links are from the IPUMS Multigenerational Longitudinal Panel. The Family Tree links are the links made by users on FamilySearch.org, and the Census Tree links are from the final Census Tree dataset.

Appendix A: Additional Figures and Tables

Appendix Figure 1: Sources on a FamilySearch Profile

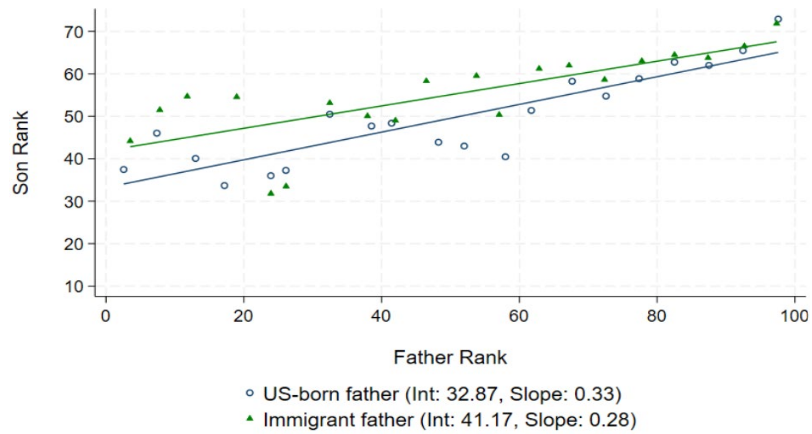


Date	Title	Created
1870	Delila A Jenkins in household of John Jenkins, "United States Census, 1870"	September 29, 2017 kimberlyhatcher3
1880	Deliah M Jinkins in household of John Jinkins, "United States Census, 1880"	September 29, 2017 kimberlyhatcher3
1900	Minnie Sharone in household of John Sharone, "United States Census, 1900"	October 11, 2017 kimberlyhatcher3
1910	Minnie Shearom, "United States Census, 1910"	June 7, 2023 ShannonBlaze
1920	Minnie Sherman in household of John Sherman, "United States Census, 1920"	June 7, 2023 ShannonBlaze
1921	Louise London in entry for Earnest Sherren and Eunice Lee Miller, "Indiana Marriages, 1811-2019"	March 13, 2020 kimberlyhatcher3
1931	Minnie Jenkins in entry for Shelby Shearons and Frances Lucile Ammons, "Kentucky, County Marriages, 1797-1954"	August 4, 2021 kimberlyhatcher3
1941	Minnie Jenkins in entry for Mr. Charlie Sherron, "Kentucky Death Records, 1911-1967"	October 25, 2021 kimberlyhatcher3
1966	Minnie Jenkins in entry for Millie Brown, "Kentucky Death Records, 1911-1967"	March 15, 2023 kimberlyhatcher3

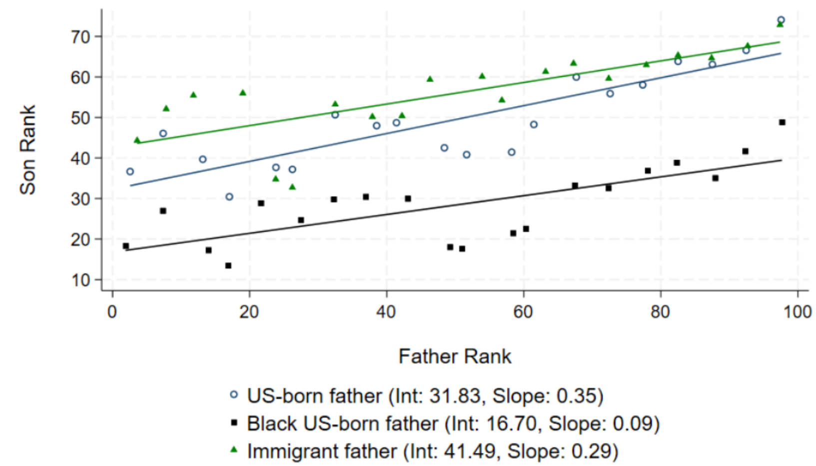
Notes: Figure shows sources attached to the profile of Delilah A. "Minnie" Jenkins, including the name of the person who attached the record. Note that Minnie's digitized name is different in each of the five attached census records.

Appendix Figure 2: Rank-Rank Estimates of Intergenerational Mobility for 1910-1940, Replicating ABJP (2021)

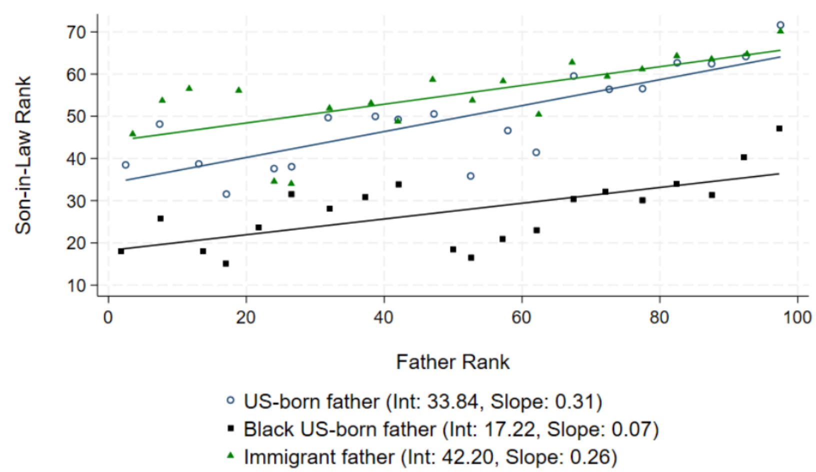
Panel A: Replication Using CLP (Men)



Panel B: Using Census Tree (Men)



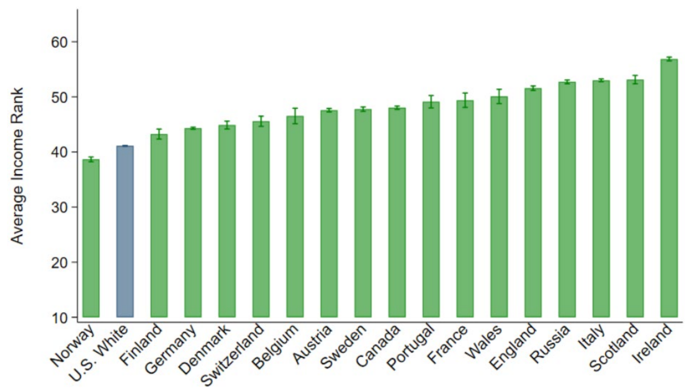
Panel C: Using Census Tree (Women)



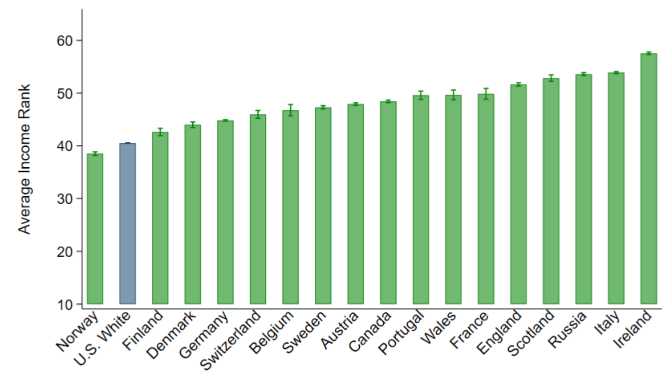
Notes: Panel A replicates the rank-rank estimates of intergenerational mobility for the 1910-1940 cohort in Figure 2 of Abramitzky, Boustan, Jácome, and Pérez (2021). Data are from the Census Linking Project. In Panel B, we produce the same estimates using the Census Tree, and add estimates for Black men. In Panel C, we use the Census Tree to produce estimates for women.

Appendix Figure 3: Average Income Rank for Children Born to 25th Percentile by Father's Birthplace, 1910-1940 Cohort, Replicating ABJP (2021)

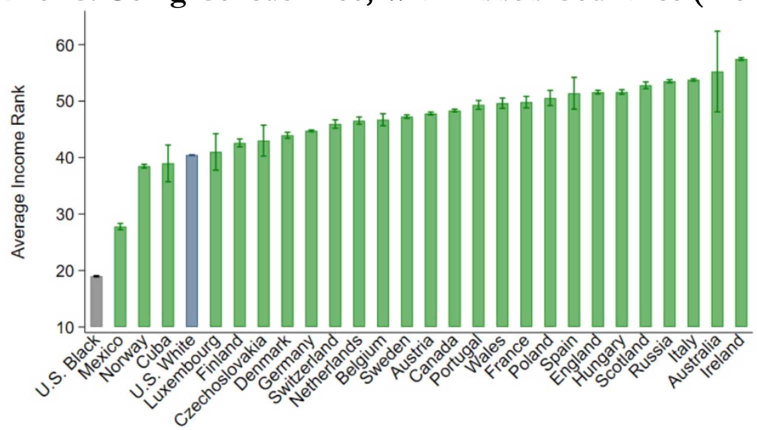
Panel A: Replication Using CLP (Men)



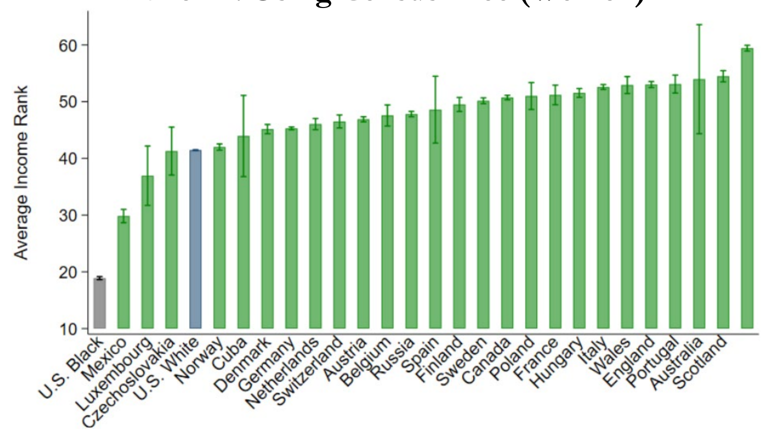
Panel B: Using Census Tree, ABJP Countries (Men)



Panel C: Using Census Tree, With Added Countries (Men)



Panel D: Using Census Tree (Women)



Notes: Panel A replicates the estimates of intergenerational mobility at the 25th percentile for the 1910-1940 cohort in Figure 3 of Abramitzky, Boustan, Jácome, and Pérez (2021). Data are from the Census Linking Project, and estimates are shown for U.S.-born Whites and for the 17 largest sending countries. In Panel B, we produce the same estimates using the Census Tree. In Panel C, we add 9 countries with samples large enough to be identified in the Census Tree. We also add estimates for Black men. In Panel D, we replicate Panel B for women. The figure shows 95% confidence intervals for each estimate.

Appendix Table 1: Features Used by XGBoost Algorithm

Category	Starting Year	Ending Year
Name		
First name JW, LV, LVN, EM	All	All
First name uniqueness interacted with JW, LV, LVN, EM	All	All
First nickname JW, LV, LVN, EM, **NYSIIS EM**	All	All
First nickname uniqueness interacted with JW, LV, LVN, EM	All	All
Middle initial EM (0 if missing)	All	All
The above feature interacted with first name EM and indicator for first initial only	All	All
Indicator for middle name longer than one letter in both years, interacted with middle name JW, LV, LVN, and EM	All	All
Last name JW, LV, LVN, EM, **NYSIIS EM**	All	All
Last name uniqueness interacted with JW, LV, LVN, EM	All	All
Birthplace		
Standardized birthplace EM	All	All
Standardized mother's and father's birthplaces EM	1880-1940	1880-1940
Standardized birthplace uniqueness	All	All
Birth year		
Absolute birth year difference <= 3	All	All
Absolute birth year difference	All	All
Sign of birth year difference	All	All
Age in starting census	All	None
Sex and marital status		
Sex EM	All	All
Female in starting census	All	All
Marital status EM	1880-1940	1880-1940
Married in starting census	1880-1940	None
Single-to-married across censuses	1880-1940	1880-1940

Appendix Table 1 (continued): Features Used by XGBoost Algorithm

Category	Starting Year	Ending Year
Household relationships		
Relationship to head EM (0 if missing)	All	All
Indicators for head, wife, son, daughter in starting census	All	All
Wife in ending census but not in starting census	All	All
Race		
Race EM	All	All
Black in starting census, and interacted with absolute birth year difference	All	All
Immigration		
Absolute immigration year difference	1900-1930	1900-1930
Indicator for immigrant in starting census	1900-1930	None
Residence		
State EM	All	All
Interaction of state EM and an indicator for not residing in birth state (0 if missing)	All	All
Township coordinate distance	All	All
Living in same place as in 1935	None	1940
Occupation		
Occupation category EM	All	All
Standardized occupation EM	All	All
Occupation string JW and EM	All	All
Raw occscore difference, and interacted with age	All	All

Notes: This table includes 66 features, of which 6 (bolded) are used for blocking. JW is Jaro-Winkler string distance. LV is Levenshtein string distance, with LVN being normalized by maximum string length. EM is exact match.

Appendix Table 2: Number of "True" Links in the Training Data Between Census Pairs

Years	Total	Women	Black
1850 to:			
1860	2,216,705	926,244	3,950
1870	1,493,049	492,216	2,488
1880	1,360,663	336,566	2,229
1900	675,397	84,119	854
1910	410,234	30,588	446
1920	172,391	9,196	171
1930	38,861	2,060	33
1940	2,757	206	4
1860 to:			
1870	2,833,051	1,201,792	5,188
1880	2,241,409	736,055	3,649
1900	1,183,510	219,844	1,392
1910	816,960	101,452	877
1920	444,347	34,165	406
1930	185,722	9,573	121
1940	43,853	2,671	34
1870 to:			
1880	4,768,988	2,028,633	51,288
1900	2,240,388	562,227	15,567
1910	1,688,831	312,420	10,148
1920	928,378	109,699	4,167

1930	572,401	43,155	2,260
1940	248,248	15,976	1,062
1880 to:			
1900	5,372,938	1,867,517	64,425
1910	4,033,175	1,030,487	36,618
1920	2,744,518	527,366	20,359
1930	1,678,316	209,626	9,654
1940	950,792	82,914	6,077
1900 to:			
1910	7,868,650	3,507,492	91,663
1920	5,707,543	1,973,705	38,706
1930	3,889,033	938,428	18,679
1940	2,543,469	432,008	12,328
1910 to:			
1920	11,687,579	5,226,222	98,074
1930	6,940,992	2,341,283	38,351
1940	4,682,932	1,171,371	22,874
1920 to:			
1930	11,728,770	5,120,313	82,500
1940	7,003,699	2,451,368	40,517
1930 to:			
1940	12,860,670	5,597,041	91,956

Notes: The XGBoost model is trained on a subset of "true" links from the Family Tree, as well as a set of "false" links.

Appendix Table 3: XGBoost Feature Importance for Adjacent Censuses

Category	Mean	1850-1860	1860-1870	1870-1880	1880-1900	1900-1910	1910-1920	1920-1930	1930-1940
Name	0.444	0.561	0.471	0.490	0.472	0.394	0.376	0.373	0.412
Residence	0.247	0.201	0.288	0.287	0.204	0.232	0.257	0.240	0.265
Birth year	0.145	0.106	0.114	0.124	0.144	0.156	0.138	0.189	0.187
Household relationships	0.065	0.083	0.087	0.063	0.078	0.058	0.061	0.055	0.030
Birthplace	0.038	0.007	0.010	0.011	0.044	0.068	0.063	0.055	0.042
Occupation	0.026	0.031	0.022	0.020	0.016	0.017	0.029	0.024	0.048
Sex and marital status	0.020	0.003	0.003	0.003	0.038	0.032	0.036	0.030	0.012
Immigration	0.014	0.000	0.000	0.000	0.000	0.042	0.038	0.032	0.000
Race	0.003	0.007	0.003	0.002	0.004	0.002	0.003	0.002	0.003

Notes: There are 66 features in the model, and here we have grouped them into categories. Blocking variables have zero feature importance; these include first name NYSIIS exact match, last name NYSIIS exact match, standardized birthplace exact match, absolute birth year difference within 3, sex exact match, and race exact match.

Appendix Table 4: Match Rates for Each Census Pair in the Census Tree

Panel A: Men

	1850	1860	1870	1880	1900	1910	1920	1930
1860	0.6686							
1870	0.5656	0.6455						
1880	0.5865	0.6166	0.7217					
1900	0.6403	0.6333	0.6481	0.6808				
1910	0.7030	0.6688	0.6585	0.6616	0.7406			
1920	0.8371	0.7471	0.7052	0.6872	0.7057	0.7904		
1930	1.1291	0.8825	0.7755	0.7302	0.7033	0.7468	0.8042	
1940	1.7628	1.1424	0.8915	0.7751	0.7189	0.7483	0.7768	0.8527

Panel B: Women

	1850	1860	1870	1880	1900	1910	1920	1930
1860	0.5922							
1870	0.4441	0.5450						
1880	0.4569	0.4975	0.6277					
1900	0.4596	0.4743	0.5075	0.5861				
1910	0.4820	0.4794	0.4951	0.5523	0.7189			
1920	0.5447	0.4966	0.4948	0.5411	0.6200	0.7427		
1930	0.7041	0.5491	0.4920	0.5255	0.5690	0.6166	0.7223	
1940	1.0800	0.6846	0.5340	0.5213	0.5519	0.5693	0.6008	0.7381

Notes: Match rates in the table are constructed as the number of links between the two years, divided by the number of people age 11 and older in the latter year, with adjustment for rates of under-enumeration in the earlier census.

Appendix Table 5: Representativeness of Census Tree, for Adjacent Censuses

	1850-1860		1860-1870		1870-1880		1880-1900	
	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)
Female	0.455	0.485	0.456	0.495	0.458	0.491	0.445	0.489
Age	31.33	30.81	32.18	31.33	32.78	32.09	40.54	33.16
White	0.988	0.978	0.982	0.878	0.925	0.877	0.933	0.887
Black	0.011	0.018	0.018	0.119	0.074	0.120	0.067	0.109
Married	0.476	0.466	0.485	0.464	0.489	0.475	0.632	0.470
HH Head	0.294	0.279	0.302	0.283	0.305	0.285	0.417	0.287
Family Members in HI	5.93	5.21	5.60	5.02	5.87	5.46	4.72	4.77
Lives in Birth State	0.610	0.512	0.611	0.550	0.632	0.573	0.630	0.602
Literate	0.923	0.912	0.866	0.785	0.868	0.830	0.911	0.883
High School Grad	-	-	-	-	-	-	-	-
N	10,433,051	18,975,196	14,146,855	27,098,171	20,850,344	35,469,052	23,995,041	56,082,690

Appendix Table 5 (continued): Representativeness of Census Tree, for Adjacent Censuses

	1900-1910		1910-1920		1920-1930		1930-1940	
	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)
Female	0.471	0.482	0.471	0.489	0.468	0.495	0.467	0.500
Age	34.22	33.59	34.95	34.60	35.83	35.37	37.12	36.69
White	0.925	0.892	0.932	0.899	0.936	0.902	0.934	0.905
Black	0.074	0.103	0.067	0.097	0.061	0.094	0.063	0.091
Married	0.502	0.482	0.521	0.509	0.528	0.517	0.527	0.523
HH Head	0.307	0.288	0.318	0.303	0.327	0.310	0.341	0.326
Family Members in HH	5.09	4.59	4.92	4.52	4.70	4.36	4.37	4.08
Lives in Birth State	0.667	0.591	0.662	0.600	0.664	0.610	0.683	0.656
Literate	0.943	0.915	0.959	0.937	0.968	0.954	-	-
High School Grad	-	-	-	-	-	-	0.253	0.257
N	45,772,617	69,725,595	55,001,392	80,497,032	66,037,498	96,202,610	78,373,335	108,342,194

Notes: Unweighted summary statistics for people linked between the two years in the Census Tree, compared to the linkable population (those age 11 and older) in the latter census.

Appendix B: Papers Using the Census Tree

Abramitzky, Ran, Jacob Conway, Roy Mill, and Luke Stein. 2023. "The Gendered Impacts of Perceived Skin Tone: Evidence from African-American Siblings in 1870–1940." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w31016.

Ager, Philipp and Viktor Malein. 2024. "The Long-Term Effects of Charity Nurseries: Evidence from Early 20th Century New York." European Historical Economics Society Working Paper #263. https://www.econstor.eu/handle/10419/302252

Althoff, Lukas, Harriet Brookes Gray, and Hugo Reichardt. n.d. "The Missing Link(s): Women and Intergenerational Mobility."

Aneja, Abhay, Silvia Farina, and Guo Xu. 2024. "Beyond the War: Public Service and the Transmission of Gender Norms." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w32639.

Bazzi, Samuel, Abel Brodeur, Martin Fiszbein, and Joanne Haddad. 2023. "Frontier History and Gender Norms in the United States." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w31079.

Buckles, Kasey, Joseph Price, Zachary Ward, and Haley E.B. Wilbert. 2023. "Family Trees and Falling Apples: Historical Intergenerational Mobility Estimates for Women and Men." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w31918.

Cache, Ellsworth, Ian Fillmore, Adrian Haws, and Joseph Price. 2023. "The Long-Run Effects of Parental Wealth Shocks on Children." Working Paper. https://www.aeaweb.org/conference/2024/program/paper/ry4ENk8e

Cools, Angela, Jared Grooms, Krzysztof Karbownik, Siobhan M. O'Keefe, Joseph Price, and Anthony Wray. 2024. "Birth Order in the Very Long-Run: Estimating Firstborn Premiums between 1850 and 1940." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w32407.

Ellsworth, Cache, Ian Fillmore, Adrian Haws, and Joseph Price. 2023. "The Long-Run Effects of Parental Wealth Shocks on Children." https://www.aeaweb.org/conference/2024/program/paper/ry4ENk8e.

Feigenbaum, James, and Gross, Daniel P. 2024. "Answering the Call of Automation: How the Labor Market Adjusted to Mechanizing Telephone Operation." *The Quarterly Journal of Economics* 139 (3): 1879–1939.

Feigenbaum, James, Jonas Helgertz, and Joseph Price. 2023. "Examining the Role of Training Data for Supervised Methods of Automated Record Linkage: Lessons for Best Practice in Economic History." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4669534.

Gabriel, Ryan, Adrian Haws, Amy Kate Bailey, and Joseph Price. 2023. "The Migration of Lynch Victims' Families, 1880–1930." *Demography* 60 (4): 1235–56. https://doi.org/10.1215/00703370-10881293.

Halliday, Timothy, Sumner La Croix, Joseph Price, and Jacob Van Leeuwen. 2024. "Male-Biased Sex Ratios, Marriage, and Household Composition in Early Twentieth-Century Hawai'i." *Asia-Pacific Economic History Review* 64 (1): 113–41. https://doi.org/10.1111/aehr.12280.

Haws, Adrian, David R. Just, and Joseph Price. 2024. "Who (Actually) Gets the Farm? Intergenerational Farm Succession in the United States." *American Journal of Agricultural Economics*. https://doi.org/10.1111/ajae.12493

Howard, Greg, and Russell Weinstein. 2022. "'Workhorses of Opportunity': Regional Universities Increase Local Social Mobility." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4244417.

Jones, Todd R., Matthew J. Millington, and Joseph Price. 2023. "Changes in Parental Gender Preference in the USA: Evidence from 1850 to 2019." *Journal of Population Economics* 36 (4): 3057–70. https://doi.org/10.1007/s00148-023-00957-x.

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

Minyo, Trevor. 2024. "The Impact of Early Youth Agricultural Education on Workforce Outcomes." Miami University. https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear=10&p10_accession_num=miami1714163836563002.

Otterstrom, Samuel M., Joseph P. Price, and Jacob Van Leeuwen. 2022. "Using Linked Census Records to Study Shrinking Cities in the United States from 1900 to 1940." *The Professional Geographer* 74 (1): 88–101. https://doi.org/10.1080/00330124.2021.1952882.

Price, Joseph, Christian vom Lehn, and Riley Wilson. 2020. "The Winners and Losers of Immigration: Evidence from Linked Historical Data." Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w27156.

Vidart, Daniela. 2024. "Revisiting the Link Between Electrification and Fertility: Evidence from the Early 20th Century United States." University of Connecticut Department of Economics Working Paper Series No. 2024-03. https://media.economics.uconn.edu/working/2024-03.pdf

Yang, Dongkyu. 2024. "Time to Accumulate: The Great Migration and the Rise of the American South." Working Paper. https://dongkyu91.github.io/jmp/Dongkyu_Yang_JMP.pdf