Subgroup Disparities in Automated Census Record Linkage

Hannah Postel (Duke)
Ran Abramitzky (Stanford and NBER)
Leah Boustan (Princeton and NBER)
Katherine Eriksson (UC-Davis and NBER)

Abstract

The automated linking of historical census records allows researchers to answer important socioeconomic questions about intergenerational mobility, immigrant assimilation, and more. We often assess automated linking methods by analyzing match rates and accuracy of the full census population. These measures vary substantially by population subgroup, such as for racial minorities and immigrant groups. This paper aims to understand why certain subgroups are more difficult to link than others and provides solutions to increase match rates and accuracy. Improving record linkage for immigrants and racial minority groups will both reduce prior bias in overall linked samples and enable study of group-specific experiences that were previously undetectable.

Introduction

Historical record linkage entails finding the same individual in two or more datasets (typically censuses) based on characteristics such as names, birthplaces, and ages. Creating linked panel data improves our understanding of important social topics like immigrant integration, intergenerational mobility, and demographic behavior. Over the past decade, multiple automated algorithms have been developed to maximize the *accuracy*, *efficiency*, *representativeness,* and *feasibility* of such efforts (Abramitzky et al 2021). Prevailing record linkage techniques generally link between 25% and 50% of the overall population; however, both match rates and accuracy are significantly lower for some population subgroups.

For example, the Abramitzky, Boustan, and Eriksson (ABE) standard linkage algorithm links 27% of native-born white men from 1900 to 1910. This match rate is similar for immigrants born in England (25%) but falls to 13% for other European nationalities (Ireland, Sweden). Only 7% of Chinese immigrants and 6% of Italian immigrants are linked across the two years. Link quality is also often lower for population subgroups. For example, manually reviewing a random sample of matches identified no implausible links for native-born whites; however, 4% of Italian matches and 16% of Chinese matches appeared incorrect. And though different algorithms tend to match the same people in the overall population (on average 96% of matches across algorithms are identical), matches are more discordant for some subgroups (e.g. only 73% of Chinese immigrant matches are identical on average). This suggests that different algorithms are not capturing the same people, which may raise concern for users of linked datasets.

Baseline match rates

Why might population subgroups be more difficult to link? In the case of immigrants, return migration may play a role. In the early 20th century, approximately one third of immigrants to the United States returned to their countries of origin (Abramitzky, Boustan, and Eriksson 2019;

Bandiera, Rasul, and Viarengo 2013). Racial minority groups (both immigrants and Black Americans) may also experience higher mortality than native-born whites, for example through employment in higher-risk jobs or greater exposure to environmental health risks (see e.g. Bakhtiari 2022). Both factors would reduce the likelihood of finding members of population subgroups across multiple census waves.

To assess whether expected match rates differ across groups, we first ask whether population subgroups experience a lower match rate "ceiling" than the native white population. That is, do we observe greater overall attrition for these groups between census years? To do so, we measure what percentage of the total male population in the United States in 1900 was still present in 1910. We employ ten-year age cohorts to exclude any children born in the intervening years, and also exclude any immigrants who reported entering the U.S. after 1900. The difference in the population size therefore represents natural decrease due to mortality and emigration. 89.7% of the native-born white population in the U.S. in 1900 was still present in 1910. This figure is lower for all subgroups; some experience slightly lower rates (~84% for African Americans and Swedish immigrants) and some diverge significantly (67% for Irish; 58% for Chinese immigrants). This exercise shows that most subgroups did experience higher mortality and/or emigration rates, which at least partially explains lower match rates. For example, when scaled according to the findings from this exercise, the Chinese match rate increases from 7% to 12%.

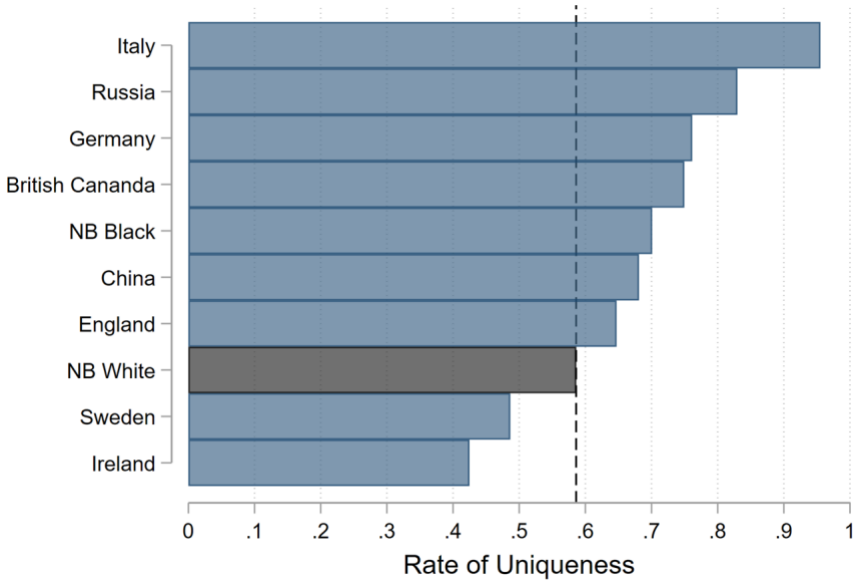
However, natural decrease does not explain the entire gap in match rates between native white populations and racial/ethnic subgroups. We suggest that a number of algorithmic and enumeration challenges may also play a significant role. First, transcription and enumeration errors for immigrant names may be higher than for Westernized names. Second, commonly used phonetic and string-distance algorithms used by record linkage algorithms to clean names were designed for use with English and may exacerbate problems for names from other linguistic origins (see e.g. Postel 2023). Finally, the lack of birthplace granularity for immigrant groups (individuals' countries of birth were enumerated, rather than state of birth for American-born individuals) makes it harder to identify matches.

This paper identifies issues and solutions for a few typologies of population subgroups. For example, Irish and Swedish immigrants exhibit low levels of uniqueness; adding extra information "breaks ties" to enable more matches. And conducting specialized pre-processing for Italian and Chinese immigrant names raises both match rates and accuracy for these groups. The following sections examine each recommendation in turn.

Recommendation 1: Add extra information

Record linkage algorithms rely on the ability to match one single individual across years. If multiple individuals are identical within a single dataset, they are excluded from consideration since it is impossible to distinguish them from others that are identical on observable characteristics. The uniqueness of a population (based on names and ages) is therefore central to the ability to link individuals across datasets. The graph below shows that Swedish and Irish immigrants are highly non-unique; over half of the population shares their name and age with at

least one other individual. This means that more than 50% of immigrants from these places are not even considered to be matched across datasets.



We therefore attempt to "break ties" by adding extra information during the match process to distinguish individuals who are not unique based on names and ages alone. Specifically, that means incorporating data on an individual's race, residence location, parental nativity, and family members' names, among other characteristics. When these extra variables are included, match rates for Irish and Swedish immigrants more than double, from ~13% to ~29%. Adding extra information helps adjudicate among individuals who previously looked identical.

Recommendation 2: Conduct specialized name cleaning

Lower match rates for other immigrant groups likely stem from issues with name enumeration and transcription. These issues include higher levels of error in how names were written down both on the original census forms and in the digitized census data, how matching algorithms handle multi-part names, and the fact that some immigrants tend to change (or "Americanize") their names.

This project draws on a recent study showing the promise of linguistically tailoring name cleaning procedures to improve match rates and accuracy. Postel (2023) more than doubled match rates for Chinese immigrants through a three-step approach addressing Chinese-specific issues in name enumeration and linkage. This paper takes a similar approach for Italian immigrants and individuals of Hispanic origin.

Italian immigrant names suffer from high levels of mis-transcription. As shown in the graph above, the Italian population is approximately 95% unique. The Italian population in 1910 (886,000 individuals) is enumerated as having over 104,000 unique first names and over 351,000 unique last names. But there is a high amount of likely misspelling. For example, 86 first names are enumerated as being only one character different from the name "Pasquale"; this affects 393

individuals.[1] If names are misspelled, they cannot be matched across years. This paper aims to harmonize and consolidate variant spellings to increase the number of potential matches across census years. We also conduct an analysis to identify potential name changes due to Americanizations (e.g. "Giuseppe" to "Joseph"); a pilot run yielded an approximately 5% increase in the Italian immigrant match rate.

Concordance

We are also interested in whether different algorithms make the same matches. We define "overlapping" matches as those where two algorithms link person 1 in Year 1 to an individual in Year 2. "Identical" matches are those where two algorithms both link person 1 in Year 1 to Person 2 in Year 2. A high share of *identical* matches of all *overlapping* matches reassures us that the matches are more likely accurate since they are made by multiple algorithms. If matches are *overlapping* but *non-identical*, at least one algorithm has made an incorrect match.

For the total population, algorithms have high levels of concordance. For example, an average of 96% of native white links are identical across algorithms. This suggests that algorithms are performing roughly similarly and identifying largely the same matches. However, concordance is lower for some subgroups, particularly Chinese, Italian, and Russian immigrants. In these cases, only 60-80% of overlapping matches are identical. This urges caution for the user, as different algorithms produce relatively dissimilar linked samples. The final version of this paper will provide a set of detailed recommendations for researchers interested in studying major population subgroups.

Disclosure Statement

The authors declare they have no relevant or material financial interests that relate to the research described in this paper, nor any research funding to report.

[1] The Levenshtein edit distance is 1, meaning only one insertion, deletion, or substitution is required to transform one string into another.