Explorations in Economic History 87 (2023) 101493

Contents lists available at ScienceDirect

Explorations in Economic History

journal homepage: www.elsevier.com/locate/eeh

Methods Article

Record linkage for character-based surnames: Evidence from chinese exclusion☆

Hannah M. Postel*

Department of Economics and Immigration Policy Lab, Stanford University, USA

A B S T R A C T

This paper proposes a novel pre-processing technique to improve record linkage for historical Chinese populations. Current matching approaches are relatively ineffective due to Chinese-specific naming conventions and enumeration errors. This paper develops a three-step process that both triples the match rate over baseline and improves match accuracy. The procedures developed in this paper can be applied in part or in full to other sources of historical data, and/or modified for use with other character-based languages such as Japanese. More broadly, this approach suggests the promise of language-specific linkage procedures to boost match rates for ethnic minority groups.

1. Introduction

Historical record linkage entails finding the same individual in two or more datasets (typically censuses) using characteristics such as names, birthplaces, and ages. Multiple automated algorithms have been developed to maximize the accuracy and efficiency of such efforts. These methods generally link between 25% and 50% of the overall population but are not equipped to handle the especially poor and uneven name transcription present for foreign-born individuals from non-English speaking countries (Abramitzky et al., 2021; Helgertz et al., 2022). These issues are particularly grave for non-Romanized names.

This project addresses the challenge of low match rates for Chinese immigrants in historical census data. A standard matching approach finds only 3.6% of Chinese men living in the United States in 1880 again in 1900, compared to 10.3% of Germans, 11.2% of Irish, and 16.3% of English.[1] These low match rates both bias overall linked samples and preclude learning about uniquely Chinese experiences in economic history. For example, Chinese immigrants in the United States were particularly central to debates around federal immigration control in the late 19th and early 20th centuries. The ability to link this group over time would both illuminate how immigration restrictions impacted Chinese communities specifically and contribute to a broader understanding of immigration policy effects.

☆ The author thanks Leah Boustan, Peter Catron, Jonas Helgertz, Brian Kernighan, Santiago Pérez, Joe Price, Sebastian Saling, Brandon Stewart, the Stewart Lab, participants in the *Explorations in Economic History* Methodological Advances in the Extraction and Analysis of Historical Data conference, and one anonymous reviewer for insightful comments. Joshua Seufert of the Princeton East Asian Library helped me find relevant Chinese-language sources, and my knowledge of basic Chinese sociolinguistics is due to Hang Du. Ken Lunde and John Jenkins kindly answered questions about the Unihan database; Myera Rashid and Harriet Brookes Gray provided helpful assistance with the Census Linking Project code and data. This project was generously supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number 1656466, the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879, the Economic History Association, the Princeton University Department of Economics Industrial Relations Section, the Princeton University Department of History, and the Princeton University Data-Driven Social Science Initiative.

* Corresponding author.
E-mail address: hpostel@stanford.edu

[1] Calculated from Census Linking Project crosswalks downloaded in June 2021 using the Abramitzky, Boustan, and Eriksson (ABE) standard algorithm with exact names (Abramitzky et al., 2012). This paper relies mainly on the ABE standard algorithm for ease of interpretation, but the proposed process can also be used with other algorithms.

https://doi.org/10.1016/j.eeh.2022.101493
Received 11 April 2022; Received in revised form 3 November 2022; Accepted 14 November 2022
Available online 25 November 2022

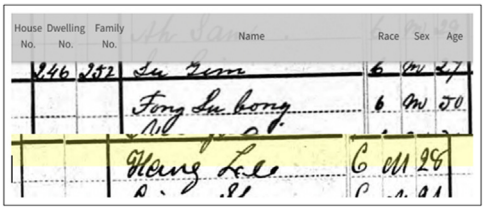

Fig. 1. Chinese names as enumerated in the 1880 census.

Table 1
Chinese names as indexed.

First Name	Last Name
—	Fong Lu Cong
Hang Lee	—

I develop a three-step pre-processing approach to increase successful record linkage for immigrant communities with character-based names in the context of late 19th-century Chinese Exclusion laws. My procedure triples the match rate over a standard approach, bringing the share of linked Chinese level with that for European immigrant groups. This project focuses on linking Chinese-born individuals between the 1880 and 1900 censuses but can also be applied to other historical data sources.

The matching approach outlined in this paper is highly adaptable. It can be used in tandem with any matching algorithm, implemented in full or in part, and customized for use depending on the available data. It can also be modified for use with other character-based languages (e.g. Japanese) to enable quantitative analysis of other historically understudied communities. Most foundationally, it highlights both the necessity and the promise of linguistically tailored matching approaches for immigrant groups. Both the code and the dictionary crosswalk are publicly available on Dataverse (Postel, 2022).

The paper proceeds as follows. Section 2 outlines three major factors influencing low match rates for Chinese immigrants. Section 3 details the steps taken to address these challenges and presents the resulting linked data. Section 4 assesses the robustness and accuracy of the proposed approach ("Postel method") in another year pair and by comparison to standard matching algorithms. Section 5 concludes.

2. Three factors contributing to low match rates

Chinese is a logographic language comprised of non-phonetic symbols ("characters"). Words (including names) can be composed of multiple characters. Contrary to many Western naming conventions, in Chinese the surname traditionally precedes the given name.[2] Though the Beijing dialect (Mandarin) is now accepted as the modern Chinese language standard, until the 1920s a large number of mutually unintelligible dialects were spoken across the country. And not until 1958 did the Chinese government institute a systematic method for transliterating characters into the Roman alphabet (*hànyǔ pīnyīn*, "spelled sounds of the spoken language of the Han people") (Chen, 1999).

U.S. census enumerators were generally unaware of these linguistic specificities. Though relatively few details on historical enumeration procedures exist, as of 1950 enumerators were instructed to "enter first the surname, then the given name in full, and the initial of the middle name, if any" (Census, 1950). It is likely that enumerators recorded names either exactly as they heard them or with oral spelling assistance from the respondent.[3]

The combination of these factors produced widespread errors in three realms of Chinese name enumeration: segmentation, name order, and standardization. Many of these enumeration issues were compounded when census pages were indexed (digitized for entry into database format). The following sections explore each challenge in turn.

2.1. Segmentation

The fact that Chinese words can – but do not always – consist of multiple individual characters poses a challenge when "translating" between Chinese characters and Romanized spellings. Word segmentation is therefore often the first step in Chinese natural language processing tasks (Chang et al., 2008). This issue is particularly evident when looking at how Chinese names were indexed in full-count census databases. For example, in one Utah county, two individuals were enumerated and indexed as follows.

Comparing Fig. 1 to Table 1 shows that these names were erroneously indexed. Name fragments were all kept together in a single column rather than being allocated across "first" and "last" name columns. Though this issue is relatively unique to individuals with

[2] See Appendix A.1 for a dictionary of frequently used terms.

[3] It is illustrative that written instructions specifying the mode of recording names do not exist in prior years. This suggests that earlier years likely suffered from less standardized approaches.

Chinese and other character-based names (as allocating e.g. "John" and "Smith" to the correct name columns was likely straightforward), for these populations such errors were widespread. Fully 79% of Chinese individuals in the 1880 census data are missing either a first or last name as indexed.

These segmentation errors pose massive challenges to record linkage, as matching algorithms compare potential matches' first names to first names and last names to last names. It is therefore extremely difficult to match individuals who are missing data in either column. This challenge is compounded by the fact that *which* name column is missing data was inconsistent even within small geographic designations. And the likelihood of *consistent* mis-segmentation across time is low, given that segmentation issues are less common in later census years.

2.2. Name order

Historical census enumerators were unsure about the order in which to record Chinese names; in fact, "it is only when Chinese characters are available that you can be sure of the position of the surname in a Chinese personal name" (Louie, 2008, 86). This led to enumerators "accustomed to European names" recording Chinese names in many "odd ways" (Jones, 1997, 31). It is therefore unclear in which order Chinese names were enumerated (i.e. according to Chinese or Anglo conventions), and unlikely name ordering practices were consistent across time or space.

2.3. Standardization

Incorrect name transcriptions occur throughout the census, particularly for recent immigrants whose names "were sometimes garbled by enumerators who had difficulty understanding foreign accents" (Archives, 2021). For Chinese immigrants, these difficulties were compounded by a lack of standardized procedure for either pronouncing Chinese characters or transliterating them to the Roman alphabet.

As mentioned above, prior to 1932 there was no standardized form of spoken Chinese. This linguistic diversity was particularly prevalent among Chinese immigrants to the United States, most of who came from southeastern Chinese provinces where multiple dialects were spoken including Cantonese (Yue), Hakka, and Min (Lai, 1991). How names were recorded by census enumerators therefore likely reflected different pronunciations across dialects.

The lack of a unified method to transliterate Chinese characters into the Roman alphabet also meant that a single character could be transliterated in many different ways. For example, the surname

A 1991 report for National Archives researchers provides census-specific validation of the above challenges (emphasis added):

> "There were also unintentional cultural barriers that stood between Chinese and the census process. For example, immigration documents bearing Chinese characters have revealed *instances of the same Chinese surname transliterated as Jung, Chung and Chong, depending on the pronunciation of the Chinese and the whim of whoever was recording it.* Moreover, another surname which was unrelated to the first was also found to be spelled Jung.
>
> This is only one example of many potential obfuscations. There was no standard method to transliterate Chinese, and census takers often wrote what they believed they heard, particularly if the Chinese could not write English well. The result is that *members of the same family or even the same individual at different times could be recorded with different transliterations of their surnames*, while members of different families could be recorded with the same transliterations" (McGlinn, 1991, 119).

This description reaffirms that the lack of a systematic Chinese transliteration scheme in combination with enumerator- and respondent-side issues yielded artificially inflated variation in census transcriptions of Chinese names.

2.4. Comparison with existing name-cleaning approaches

Low match rates are perhaps not surprising in view of these many challenges: the underlying data quality for historical Chinese immigrants in the U.S. Census is much lower than for other groups. But modern research approaches (in the form of record linkage algorithms) also contribute to low linkage rates by mishandling Chinese names.

Most fundamentally, by relying on linking first names to first names and last names to last names as indexed, these approaches are highly vulnerable to errors in segmentation and name ordering. A drastic example occurs in how the Abramitzky, Boustan, and Eriksson (ABE) algorithm's built-in name cleaning procedure handles multi-part names. In most cases, names with more than one fragment are cleaned to drop all but first fragment. This means that the 1625 multi-part raw names including "Chin" (e.g. "Chin Fung", "Chin Hing", "Chin Lung") are all condensed to simply "Chin" after cleaning. This not only over-condenses multiple-part names, but also makes them equivalent to the 417 individuals listed with only "Chin" as a first name. This issue is particularly dire given the widespread name segmentation issues in the 1880 census.

Existing record linkage algorithms also all involve some form of name standardization, employed to harmonize multiple variations of the same person's name (through e.g. remedying inconsistent spellings, standardizing abbreviations, and applying nickname equivalents). Unfortunately, these procedures – both phonetic and string-based – do not work well in Chinese. ABE employs a phonetic similarity algorithm (NYSIIS) that tends to over-condense Chinese surnames (Li et al., 2018). For example, the distinct surnames "Yang", "Ying", and "Young" are all standardized as "yang". Two other approaches (the Multigenerational Longitudinal Panel (MLP) and BYU Census Tree) employ string distance measures, which often designate two Chinese names as different when they in fact share the same underlying character (Peng et al., 2015).

Table 2
Segmenting multi-part names into individual fragments.

As transcribed		After name splitting			
First name	Last name	Name1	Name2	Name3	Name4
—	Fong Lu Cong	Fong	Lu	Cong	.
Hang Lee	—	Hang	Lee	.	.

The procedure outlined below addresses each of the above issues to boost both match rate and match accuracy for historical Chinese populations. It links 7345 individuals between the 1880 and 1900 censuses (9.6% of the 1900 population), performing similarly well in a different year pair (see Section 4). This represents a three-fold increase over a standard baseline match. All data and code are available online (Postel, 2022).

3. Matching approach and results

This paper develops a three-step pre-processing approach for character-based names to be iteratively applied with the researcher's choice of record linkage algorithm. For the sake of clarity, this paper uses only the fully automated ABE standard matching approach but can be used – in part or in full – with any other algorithm.

I use the 1880 and 1900 full-count census data provided by IPUMS and accessed via the National Bureau of Economic Research.[4] My data consist of men born in China and designated to be of Chinese race living in the continental United States: 97,970 in 1880 and 76,484 in 1900.[5] The significant population decrease can be attributed to the restrictive and increasingly stringent Chinese Exclusion laws implemented during this period (see e.g. Salyer, 1995).

I conduct a baseline match following basic name cleaning procedures including replacing special characters, trimming spaces, and standardizing common name spellings.[6] 2418 individuals are linked at baseline for a match rate of 3.2%.[7]

3.1. Segmentation

The first step of this paper's proposed approach addresses segmentation errors by consistently delimiting name fragments within their indexed first- and last-name columns. I ensure each name fragment (here conceptualized as single Chinese character) is separated by a space within the indexed column, next segmenting each fragment into its own column (shown in Table 2 below).[8] In the example of "Fong Lu Cong" from Utah above, "Fong", "Lu", and "Cong" are each individual fragments.

A standard ABE match conducted following this segmentation step compares each name fragment among potential matches (e.g. "Name 1" with "Name 1", "Name 2" with "Name 2", etc). This step links 2225 additional individuals – a 92% increase over baseline – for a total of 4643 linked pairs.

3.2. Name order

Step 2 attempts to remedy inconsistent name order enumeration. I do so by changing the order of the name fragments created in the previous step for one year's data only. Specifically, this step matches the original 1880 name fragments to a 1900 dataset with name fragments listed in a different order.

I conduct this approach separately for two- and three-fragment names.[9] In the simple case of a two-fragment name, "Hang Lee" would be left identical to the previous step in 1880 and reversed to "Lee Hang" in 1900. I change the order for three-fragment names in two ways reflecting the fact that most Chinese surnames consist of only one character.[10] Suppose an individual in 1900 was enumerated as Ping Deng Xiao. The accurate surname is likely listed as either fragment one ("Peng") or fragment three ("Xiao"), since a single-character family name would be enumerated either first or last. I therefore try matching after converting this individual's name first to Xiao Ping Deng and next to Deng Xiao Ping (which we know to be the correct order in the case of a former Chinese leader).

The step links an additional 2193 individuals for a running total of 6836 matched pairs.

[4] Section 5 replicates the process for 1900–1910.

[5] Though Hawaii was enumerated in the 1900 census, I do not include it here as it was not yet a state and subject to separate immigration laws (Glick, 1980).

[6] I do not use the built-in 'abeclean' command due to the issues with name segmentation and standardization discussed above.

[7] This paper calculates the match rate out of the 1900 population given 1) the relative enumeration quality and 2) for comparison with the 1900–1910 matches in Section 4.

[8] Except for two-character surnames (Ouyang and Situ), the only two-part surnames "brought by early immigrants" (Louie, 2008, 119). I treat both as a single fragment, condensing them into a single column so they aren't interpreted as non-surname fragments.

[9] The large majority of individuals had two-fragment names. Only 3% of the population in 1880 had a three- or four-part name, though this increased to 12% in 1900. I adjudicate potential matches for four-part names manually given their scarcity (just 18 in 1880 and 154 in 1900).

[10] See footnote 8 for more information on two-character surnames.

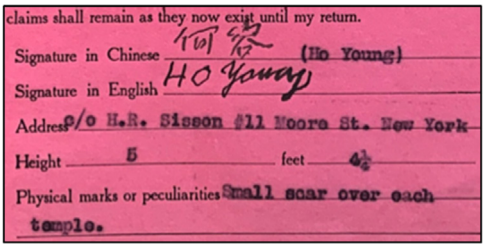
claims shall remain as they now exist until my return.
Signature in Chinese (Ho Young)
Signature in English HO Young
Address c/o H.R. Sisson #11 Moore St. New York
Height 5 feet 4½
Physical marks or peculiarities Small scar over each temple.

Fig. 2. Ho Young Exclusion file.

Table 3
Excerpt from character-equivalent crosswalk.

Name1	N1_freq	C1_utf	C1_py	C1_freq	C1_prop	C2_utf	C2_py	C2_freq	C2_prop
Ho	49	U+4F55	hé	49	1.00	.	.	.	.
Fong	137	U+65B9	fāng	67	.489	U+913A	zhào	31	.229

Notes. The column N1_freq lists the total number of times a specific Romanization appeared in the New York Exclusion File index; C*_freq the number of times a character was mapped to said Romanization. C*_prop calculates the frequency with which a character maps to a Romanization; C*_utf and C*_py list the character in its UTF-8 and *pinyin* forms, respectively.

3.3. Standardization

Finally, I conduct name standardization customized for Chinese names. The goal of this step is to minimize artificially inflated variation in indexed names by mapping Romanized spellings to their equivalent Chinese character. This is essentially a form of dimension reduction based on the linguistic traits described in Section 2.3 above.

I draw on a historical database of Chinese immigration files to compile a standardization crosswalk consisting of equivalencies between Romanizations and characters. This step focuses on surnames, as the number of potential character mappings is at least an order of magnitude smaller.[11]

3.3.1. Creating a crosswalk

My historical data source is a digitized index of Chinese Exclusion case files covering the New York region. Chinese immigrants to the United States had to complete and sign many application forms upon entry and exit; these signatures were often recorded in both Romanized and Chinese character formats, as shown in Fig. 2 below.

This index – accessed at the U.S. National Archives' New York City branch – consists of 18,533 individuals. 60% are listed with surnames in both Romanized and character forms. Following a validation process described in Appendix A.2, I merge the Exclusion File index with the Unihan database, which "contains mapping data to allow conversion to and from other coded character sets" based on Chinese ideographic script.[12]

The final product is a crosswalk of character mappings for each Romanized name. I calculate the total number of each Romanized surname in the New York database and the frequency of each identified character mapping. Table 3 below gives two examples.

The surname "Ho" appears 49 times in the index and each time is identified as corresponding to the character

3.3.2. Applying crosswalk to census data

The next step is to apply the name crosswalk to the census data. In some cases this is straightforward, such as for surname "Ho". For cases such as "Fong", however, there is no dispositive character equivalent as even the most common character mapping occurs for less than half of cases.

I propose using the listed frequencies to inform a decision rule on when to apply a character equivalent to a Romanized name. Given that some Romanizations match to more than one character, the mapping frequencies can serve as a threshold over which the user "trusts" the character mappings of choice. This analysis specifies a cutoff of 0.7 but alternative thresholds could be employed.[13] After applying the crosswalk to each census year and replacing relevant names with their character equivalent (in UTF-8 format), an additional matching attempt is made. This step identifies an additional 509 matched pairs.[14]

[11] An 1892 publication listed approximately 2000 surnames in use in China, compared to tens of thousands of options for personal names. And a study of late nineteenth-century Chinese civil servants found that 70% shared the same 55 surnames (Giles, 1892).

[12] This resource was compiled by the Unicode Consortium and is freely available on the Unicode website: https://www.unicode.org/reports/tr38/.

[13] For example, specifying a cutoff of 0.5 yields 850 new matches, while using 0.9 yields 325. The decision rule could also be multipronged and/or probabilistic, incorporating total proportion, total frequency, and relative frequency compared to the next most common character.

[14] Matching is conducted in both the original and reversed name order as described in Section 3.2 above.

Table 4
Balance Test of Total Population vs Matched Observations, 1880–1900.

	1880 pop.	1880 Postel	1880 ABE	1900 pop.	1900 Postel	1900 ABE
Age	31.8	-4.45	-4.59	41.9	6.3	5.9
	(10.7)	(0.11)	(0.16)	(10.8)	(0.11)	(0.17)
Single (i.)	0.69	0.05	0.06	0.58	-0.02	-0.06
	(0.46)	(0.01)	(0.01)	(0.49)	(0.01)	(0.01)
Literate (i.)	0.77	0.03	0.01	0.72	-0.04	-0.03
	(0.42)	(0.01)	(0.01)	(0.45)	(0.01)	(0.01)
Socioeconomic	13.7	0.85	1.01	16.3	-1.69	-1.32
Index	(13.4)	(0.16)	(0.25)	(17.6)	(0.20)	(0.33)
Lives in CA (i.)	0.70	-0.04	-0.01	0.50	0.09	0.04
	(0.46)	(0.01)	(0.01)	(0.50)	(0.01)	(0.01)

Notes. The columns marked "pop" correspond to the population mean and standard deviation of all observations for the specified year. An "i." designates an indicator variable; reported figures correspond to the share of the population with that characteristic. The table presents the balance test between my approach and the ABE standard algorithm. Estimates correspond to the difference in averages between the total and matched observations for a given algorithm and are reported in the first line; standard errors are reported in parentheses in the corresponding second line.

Table 5
Match rate comparisons.

	1880–1900	1900–1910
Postel	7,345	7,272
	(9.6%)	(9.5%)
ABE	2,721	5,566
	(3.6%)	(7.3%)
MLP	46	2,160
	(0.06%)	(2.8%)
Census Tree	———	12,855
		(16.8%)

3.4. Results

The Postel method described in this paper links 7345 Chinese individuals between 1880 and 1900. A three-fold increase over baseline, these gains were driven mostly by segmentation and name order changes (accounting for 45.2% and 44.5% of new matches, respectively). Character standardization contributed the remaining 10.3%. I link 9.6% of the 1900 Chinese immigrant population, on par with match rates for European immigrant groups.

I identify nearly three times as many matches as the ABE exact standard approach for 1880–1900 (2721 or 3.6%). The IPUMS Multigenerational Longitudinal Panel (MLP) links less than one tenth of one percent of individuals (just 46) between these years.

Table 4 compares the observable characteristics of matched individuals with the overall population mean for each year. The Postel method and the ABE standard approach yield similarly representative samples, though both demonstrate some bias compared to the overall Chinese population particularly in terms of age.[15] As with other linked data, projects using this matching approach should both discuss the sample's representativeness and attempt to reweight the sample based on observed characteristics (Abramitzky et al., 2021).

4. Robustness and accuracy

This section conducts a robustness exercise, applying the approach detailed above to another census year pair (1900–1910). I also compare the quantity, accuracy, and uniqueness of my links to those produced by ABE, MLP, and BYU Census Tree.[16]

4.1. Quantity

Table 5 below shows that the Postel method matches 7272 individuals from 1900 to 1910 (9.5% of the 1900 population). This result is second only to the BYU Census Tree, which links 12,855 individuals in this year pair. The Census Tree can be considered an upper-bound estimate of a semi-automated approach, given the volume of their individually validated training data (Price et al., 2021).

[15] The MLP match rate is too low to yield useful comparison.

[16] This year pair is more suited for comparison as the BYU Census Tree is not available for 1880–1900, and MLP's extremely low 1880–1900 match rate makes assessing accuracy difficult. The analysis is limited to men for comparability across all approaches.

Table 6
Match accuracy (1900–1910).

	Postel	ABE	MLP	CT
Exact	41	54	24	25
	(20.5%)	(27.0%)	(12.0%)	(12.5%)
Likely	42	48	19	32
	(21.0%)	(24.0%)	(9.5%)	(16.0%)
Plausible	117	59	62	73
	(58.5%)	(29.5%)	(31.0%)	(36.5%)
Implausible	0	39	95	70
	(0.0%)	(19.5%)	(47.5%)	(35.0%)

Table 7
Uniqueness of Postel matches, 1900–1910.

	Matched pairs (% of total)
Unique	2860 (39.3%)
Shared with one other	2303 (31.7%)
ABE[20]	1313
MLP	71
CT	919
Shared with two others	1768 (24.3%)
ABE+MLP	99
ABE+CT	1584
MLP+CT	85
Found by all	341 (4.7%)
Total Postel	**7272**

Both ABE and MLP demonstrate higher match rates than for 1880–1900, but are still significantly outstripped by the method proposed in this paper.

4.2. Accuracy

The prior section demonstrates how this approach increases the *quantity* of successful links. What – if anything – can we say about link *quality*? Unfortunately, the lack of "ground truth" data for the time period and demographic group under consideration precludes any systematic assessment of the accuracy of Postel method links.[17]

However, a manual validation exercise conducted on a random sample of links is illustrative. 200 matches from each linkage approach were randomly selected and reviewed for accuracy. Table 6 below reports the results of this exercise.

An "exact match" means both names and ages are identical between the two years. A "likely" match consists of individuals with identical names and ages one year apart (e.g. if an individual is 42 years old in 1900 and 51 or 53 in 1910 with exact name correspondence they are considered a likely match). I consider a match "plausible" if individuals either have 1) identical names and ages two years apart or 2) similar – but not identical – names with ages one year apart. Finally, "implausible matches" are matched individuals whose ages are more than three years apart or have very different names.[18]

This exercise demonstrates that Postel matches are of higher quality than those produced by other algorithms. This improvement is almost entirely due to eliminating "implausible" matches through language-specific name pre-processing.[19] The matches produced by ABE are of next highest quality, likely due the "exact" approach requiring exact name matches (i.e. no name standardization) and limiting age discrepancies to a five-year band. In contrast, the MLP and Census Tree algorithms yielded both the fewest "exact" matches and the most "implausible" matches. Likely due to these algorithms' flexibility, a high share of the randomly sampled matches suffered from large age differentials and extremely different names.

These differences in match quality are reflected in a balance test for 1900–1910 (see Appendix A1). The Postel and ABE methods generally yield more representative samples than Census Tree or MLP links.

[17] Abramitzky et al. (2021), for example, use two different 1940 census transcriptions to benchmark their match rates. Alternative "ground truth" data sources typically rely on extensive hand links, such the Union Army data compiled by the Early Indicators project (Costa et al., 2018) and the 1915 Iowa census data indexed by Feigenbaum (2016). Unfortunately, neither of these sources is well-suited to this project, as Chinese individuals could not join the U.S. military during the period under consideration and very few resided in Iowa.

[18] Names are considered similar when one name fragment varies but the others are identical; names with variations across all name fragments are considered different.

[19] I designate matches from the reversed name order step as "plausible" given the lack of a ground truth "correct" name order.

[20] The relatively high overlap with ABE makes sense given that I use their matching approach after pre-processing.

4.3. Uniqueness

Inspecting match uniqueness further underscores the Postel method's similarities and differences from other algorithms. As seen in Table 7 below, this paper's approach identifies 2860 matched pairs not linked by any other algorithm. The remaining 60.7% of matches are found by at least one other matching algorithm.

Of the matches unique to this approach, 680 stem from the baseline match, 291 from the segmentation step, 1525 from the name order step, and 364 from standardization. Only the Postel method reverses the name order of potential matches so it makes sense this step accounts for the largest number of unique matches. This comparison also highlights the importance of language-specific name cleaning procedures, given that my baseline match diverges from the other methods mainly by avoiding erroneous adjustment by phonetic or string-distance algorithms.

5. Conclusion

This project develops an approach to remedy low record linkage rates for Chinese immigrants in the 19th- and 20th-century United States. The ability to track members of the first major non-white immigrant group in the United States across time enables insights into core sociopolitical processes like racial boundary formation, immigrant integration, and immigration policy effectiveness. This approach triples the number of matches compared to baseline, bringing Chinese match rates in line with that of some European immigrant groups. Both the quantity and the quality of these matches are higher compared to standard record linkage approaches.

This approach could also be used to match Chinese women, as the convention of changing one's surname is not present in Chinese.[21] This was a small population but substantively important to understand gender-based immigration restrictions. Additionally, a similar approach could be tailored for other East Asian immigrant groups with character-based languages (e.g. Japanese).

More broadly, the procedures developed in this paper indicate the importance of group-specific name pre-processing for record linkage. Given relatively low match rates for immigrant groups, using linguistically-tailored matching approaches offers an important opportunity to study minority group experiences and reduce bias in overall linked samples.

Data availability

Data are available at https://doi.org/10.7910/DVN/LWPOJQ. Access to the underlying full-count census data is restricted; licenses are available via IPUMS.

Appendix A

A1. Commonly used terms

- "First name" and "last name" refer to an individual's name as enumerated in the indexed census database (i.e. IPUMS full-count datasets).
- "Surname" and"personal name" refer to an individual's "real" name in Chinese. As discussed above, an individual's true surname could have been recorded in either the "first name" or "last name" columns of the indexed census data.
- "Name fragments" refer to each individual part (typically single character) of a name.
- "Indexed" or "transcribed" refers to the computer-readable database (e.g. IPUMS full-count census data) while "enumerated" refers to the underlying census forms (handwritten).
- A "crosswalk" is a file mapping equivalent elements across two or more distinct datasets.

A2. Identifying character equivalents in the New York Exclusion file database

Surname characters in the New York Exclusion database were identified with four-corner codes, a now mostly obsolete codification scheme to numerically encode characters. Validating and refining the list of four-corner codes yields a crosswalk of 413 Romanizations mapped to 152 unique character equivalents. Four-corner codes narrow the possible character set immensely, but they are not always a unique character identifier.

To ensure I choose the correct character equivalent, I validate each four-corner code in the MBDG online dictionary. I choose the relevant character based on whether its definition indicates use as a surname. If the MDBG definitions do not indicate surname use, I search the genealogical website My China Roots using the Romanized spelling in the New York database and compare potential options with those offered by the MDBG four-corner code search. I leave the entry blank if no character equivalent is found. I further refine the possible character set by expanding the four-corner code by one digit. For example, the four-corner code 8211 produces eleven possible character matches. Specifying code 8211.4 narrows the choice set to two characters, only one of which is designated as a surname.

[21] I do not do so in this exercise for consistency with other match rates (e.g. as calculated in the CLP crosswalks).

Table A1
Balance Test of Total Population vs Matched Observations, 1900–1910.

	1900 pop.	Postel	ABE (std)	MLP	CT	1910 pop.	Postel	ABE (std)	MLP	CT
Age	41.9	-3.21	-2.86	-2.16	-3.21	44.9	4.66	4.30	5.33	5.52
	(10.8)	(0.18)	(0.13)	(0.20)	(0.09)	(12.9)	(0.12)	(0.15)	(0.19)	(0.11)
Single (i.)	0.58	0.04	-0.01	-0.09	0.07	0.55	-0.05	-0.07	-0.11	-0.05
	(0.49)	(0.01)	(0.01)	(0.01)	(0.01)	(0.50)	(0.01)	(0.01)	(0.01)	(0.01)
Speaks English (i.)	0.62	0.02	0.01	-0.04	0.01	0.56	0.03	0.02	-0.05	0.02
	(0.48)	(0.01)	(0.01)	(0.01)	(0.01)	(0.50)	(0.01)	(0.01)	(0.01)	(0.01)
Literate (i.)	0.72	0.02	0.03	0.06	0.03	0.82	-0.02	-0.02	0.06	-0.02
	(0.45)	(0.01)	(0.01)	(0.01)	(0.004)	(0.39)	(0.01)	(0.01)	(0.07)	(0.004)
Socioeconomic Index	16.3	0.01	0.92	4.38	0.95	21.2	0.30	1.28	3.52	1.17
	(17.6)	(0.21)	(0.25)	(0.47)	(0.17)	(22.3)	(0.28)	(0.34)	(0.54)	(0.23)
Immigration Year	1879.1	1.29	1.13	1.24	1.55	1886.8	-3.78	-3.52	-5.13	-4.21
	(8.6)	(0.10)	(0.11)	(0.16)	(0.08)	(11.8)	(0.13)	(0.15)	(0.20)	(0.11)
Lives in CA (i.)	0.50	-0.02	-0.01	0.11	0.01	0.50	0.04	0.02	0.13	0.03
	(0.50)	(0.01)	(0.01)	(0.01)	(0.01)	(0.50)	(0.01)	(0.01)	(0.01)	(0.01)

Notes. The columns marked “pop.” correspond to the population mean and standard deviation of all observations for the specified year. An “i.” designates an indicator variable; reported figures correspond to the share of the population with that characteristic. The table presents the balance test between my approach and three other standard matching algorithms. Estimates correspond to the difference in averages between the total and matched observations for a given algorithm and are reported in the first line; standard errors are reported in parentheses in the corresponding second line.

References

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., Pérez, S., et al., 2021. Automated linking of historical data. J. Econ. Lit. 59 (3), 865–918. doi:10.1257/jel.20201599.

Abramitzky, R., Boustan, L.P., Eriksson, K., et al., 2012. Europe’s tired, poor, huddled masses: self-selection and economic outcomes in the age of mass migration. Am. Econ. Rev. 102 (5), 1832–1856. doi:10.1257/aer.102.5.1832.

Archives, U. S. N., 2021. 20 Tips for Census Research Success. https://www.archives.gov/research/census/20-tips-for-census-research-success.

Census, U. S. B. o. t., 1950. Urban & Rural Enumerator’s Reference Manual, 1950 Census of the United States. Washington, D.C.

Chang, P.-C., Galley, M., Manning, C.D., et al., 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In: Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, Columbus, Ohio, pp. 224–232. https://aclanthology.org/W08-0336

Chen, P., 1999. Modern Chinese: History and sociolinguistics. Cambridge University Press.

Costa, D.L., Kahn, M.E., Roudiez, C., Wilson, S., et al., 2018. Data set from the Union Army samples to study locational choice and social networks. Data in Brief 17, 226–233. doi:10.1016/j.dib.2017.12.007. http://www.sciencedirect.com/science/article/pii/S2352340917307023

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

Giles, H.A., 1892. A Chinese-English dictionary. Kelly and Walsh.

Glick, C.E., 1980. Sojourners and settlers: Chinese migrants in Hawaii. University Press of Hawaii.

Helgertz, J., Price, J., Wellington, J., Thompson, K.J., Ruggles, S., Fitch, C.A., et al., 2022. A new strategy for linking U.S. historical censuses: a case study for the IPUMS Multigenerational Longitudinal Panel. Historic. Method. 55 (1), 12–29. doi:10.1080/01615440.2021.1985027.

Jones, R., 1997. Chinese names: The traditions surrounding the use of Chinese surnames and personal names. Pelanduk Publications, Singapore.

Lai, H.M., 1991. Guangdong Regional and Historical Background, Genealogy Research Information. Technical Report. San Francisco, California, United States.

Li, M., Danilevsky, M., Noeman, S., Li, Y., et al., 2018. DIMSIM: An Accurate Chinese Phonetic Similarity Algorithm Based on Learned High Dimensional Encoding. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. Association for Computational Linguistics, Brussels, Belgium, pp. 444–453. doi:10.18653/v1/K18-1043.

Louie, E.W., 2008. Chinese American names: Tradition and transition. McFarland.

McGlinn, L.A., 1991. Early Chinese immigrants and the United States census. Middle States Geographer 24, 113–120.

Peng, N., Yu, M., Dredze, M., et al., 2015. An Empirical Study of Chinese Name Matching and Applications. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Beijing, pp. 377–383. doi:10.3115/v1/P15-2062.

Postel, H., 2022. Replication Data for: Record Linkage for Character-Based Surnames - Evidence from Chinese Exclusion. Technical Report. Harvard Dataverse. https://doi.org/10.7910/DVN/LWPOJQ

Price, J., Buckles, K., Van Leeuwen, J., Riley, I., 2021. Combining family history and machine learning to link historical records: the Census Tree data set. Explor. Econ. Hist. 80, 101391. doi:10.1016/j.eeh.2021.101391. https://www.sciencedirect.com/science/article/pii/S0014498321000024

Salyer, L.E., 1995. Laws harsh as tigers: Chinese immigrants and the shaping of modern immigration law. Chapel Hill: University of North Carolina Press.