

Aaronson, Daniel; Davis, Jonathan; Schulze, Karl

Working Paper

Internal immigrant mobility in the early 20th century: Experimental evidence from Galveston immigrants

Working Paper, No. 2018-04

Provided in Cooperation with:

Federal Reserve Bank of Chicago

Suggested Citation: Aaronson, Daniel; Davis, Jonathan; Schulze, Karl (2018) : Internal immigrant mobility in the early 20th century: Experimental evidence from Galveston immigrants, Working Paper, No. 2018-04, Federal Reserve Bank of Chicago, Chicago, IL

This Version is available at:

<https://hdl.handle.net/10419/200574>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



Federal Reserve Bank of Chicago

**Internal Immigrant Mobility in the Early
20th Century: Experimental Evidence
from Galveston Immigrants**

*Daniel Aaronson, Jonathan Davis, and
Karl Schulze*

February 2018

WP 2018-04

**Working papers are not edited, and all opinions and errors are the responsibility of the author(s). The views expressed do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.*

Internal Immigrant Mobility in the Early 20th Century: Experimental Evidence from Galveston Immigrants

Daniel Aaronson

Federal Reserve Bank of Chicago

Jonathan Davis

University of Chicago

Karl Schulze

Federal Reserve Bank of Chicago

February 2018

Abstract

Between 1907 and 1914, the “Galveston Movement,” a philanthropic effort spearheaded by Jacob Schiff, fostered the immigration of approximately 10,000 Russian Jews through the Port of Galveston, Texas. Upon arrival, households were given train tickets to pre-selected locations west of the Mississippi River where a job awaited. Despite the program’s stated purpose to locate new Russian Jewish immigrants to the Western part of the U.S., we find that almost 90 percent of the prime age male participants ultimately moved east of the Mississippi, typically to large Northeastern and Midwestern cities. We use a standard framework for modeling location decisions to show destination assignments made cities more desirable, but this effect was overwhelmed by the attraction of religious and country of origin enclaves. By contrast, there is no economically or statistically significant effect of a place having a larger base of immigrants from other areas of the world and economic conditions appear to be of secondary importance, especially for participants near the bottom of the skill distribution. Our paper also introduces two novel adjustments for matching historical data – using an objective measure of match quality to fine tune our match scores and a deferred acceptance algorithm to avoid multiple matching.

Aaronson: daaronson@frbchi.org; Davis: jonmvdavis@gmail.com; Schulze: kschulze@frbchi.org. We thank participants at the Social Science History Association Conference and especially Phyllis Aaronson for spreading the story of her father, a Galveston immigrant who wound up in NYC. The views expressed in this paper are not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

I. Introduction

One of the more compelling but challenging questions in the social sciences is the extent to which place determines socioeconomic success. The best evidence to date stems from two major U.S. programs -- the Gautreaux assisted housing project in Chicago and its national successor, Moving to Opportunity (MTO). Each assigned low-income residents to more affluent and racially mixed neighborhoods based on quasi-random and random designs, respectively. Our paper is the first that we are aware of to study an immigration program that in many ways is a more ambitious, early precursor to these influential experiments.

Starting in the mid-19th century, a mass migration brought tens of millions of European immigrants to the U.S. (Hatton and Williamson 1994). Many disembarked at Ellis Island and then proceeded to a variety of Eastern and Midwestern urban areas. East European Jewish immigrants that were part of this wave clustered heavily in New York City. By 1907, the large, unprecedented influx into already dense parts of NYC prompted a prominent local Jewish philanthropist, Jacob Schiff, to partly fund a program that ultimately steered roughly 10,000 Russian Jews from more than 7,000 families to a new U.S. gateway – the port of Galveston, Texas. Upon arrival in Galveston, each family was given train tickets to destinations west of the Mississippi River that were determined by matching the male head’s occupation with local demand for their skill.

Since Schiff’s immigrants had little input on their initial assignment and indeed much of the decision was based primarily on a single observable characteristic, this setting in many ways mirrors later mobility experiments.¹ However, we cannot study the role of place in shaping life

¹ Similar natural experiments in Sweden and Denmark are described in Edin, Fredriksson, and Aslund (2003) and Damm and Dustmann (2014).

outcomes because, despite the stated purpose of the program, roughly 88 percent of prime-age Galveston men sent to the western U.S. wound up *east* of the Mississippi River, and especially in NYC and Chicago.

To explain this lack of compliance, we use a multinomial logit framework inspired by a Roy model for location decisions. Each individual chooses from the 48 possible U.S. states, with demographic and economic characteristics determining relative attractiveness. The program had a clear impact. Immigrants were more likely to reside in their assignment state a decade and more later; in the model, the increase in their latent utility for their assignment state is equivalent to increasing the number of local Russian immigrants, the strongest determinant of locational preference, by a factor of 20, holding all other features of the state fixed. However, this seemingly large effect resulted in only a minor shift in the ultimate location decisions of the Galveston immigrants since other Russian and Jewish immigrants were overwhelmingly concentrated in Northeastern and Midwestern states. For example, in 1910, the state of New York had 17 times more Russian immigrants than North Dakota, the state west of the Mississippi with the largest Russian immigrant population, and differed on a number of other dimensions that may have made it more attractive as well. As a result, the Galveston assignment could not overcome the appeal of living in areas with a high density of Russians or Jews. For a typical state, assignment corresponds to a highly significant 1.9 percent increase in the probability of choosing that state.

While the number of Russian immigrants is especially powerful in predicting where Galveston immigrants ultimately resided, by contrast there is no economically or statistically significant effect from a place having a larger base of immigrants from other areas of the world. Moreover, while we find an association between observable economic conditions and location,

these appear to be of secondary importance to our ethnic measures, especially for participants near the bottom of the skill distribution. In total, we interpret this evidence as consistent with ethnic and religious networks playing a vital role in location decisions for early 20th century Russian Jewish immigrants, paralleling the experiences of many other immigrant waves to the U.S. over the remainder of the 20th century (e.g. Robinson 1998; Haines 1989; Bartel 1989; Cutler, Glaeser, and Vigdor 2008; Damn 2009; and Beaman 2012). Indeed, this result has some parallel to MTO, where participants ultimately sorted into neighborhoods with less poverty but similar racial composition to their original communities (Orr et al. 2003).

The main empirical challenge we face is successfully matching administrative records from the Port of Galveston to the 100 percent population counts from the 1910 to 1940 decennial censuses. Bailey et al (2017) provide compelling evidence that hand matching administrative records is by far the gold standard methodology. In contrast, standard matching algorithms that principally rely on phonetic codes, such as Soundex or the New York State Identification and Intelligence System (NYSIIS), can have high rates of both type I and type II errors. However, in practice, hand matching is far more costly than the algorithmic approach.

We build on existing algorithmic methods (e.g. Ferrie 1996 and Abramitzky et al. 2012 but especially Feigenbaum 2016) in two novel ways. First, like Feigenbaum (2016), we tune our scoring of potential matches based on an objective measure of match quality. However, rather than relying on a training sample of hand-matches, we measure match quality using the share of Galveston immigrants who arrived after 1910 who were improperly (because they were not in the U.S.) matched to the 1910 census. Using this calibrated scoring procedure, the pre-1910 arrival match rate with the 1910 census is 73 percent whereas the post-1910 match rate is 11 percent. Moreover, we get similar match rates if instead we calibrate our procedure using half

our data and evaluate the matches using a hold-out sample. This method is a scalable approach to improving match quality between immigration records and decennial censuses. Our second nonstandard technique is to select best matches using a deferred acceptance algorithm (DAA) (Gale and Shapley 1962). The DAA allows us to keep Galveston immigrants who are matched to multiple census observations but avoid the real possibility that multiple Galveston immigrants are matched to the same individual in a Census. Nix and Qian (2015) show that observations with multiple potential census matches substantially increase the match rate. To the best of our knowledge, the potential for multiple matches on the same census record is not typically addressed in the literature (an exception is Feigenbaum 2016). Using the diagnostics developed in Bailey et al (2017), we show that our approach yields a match rate nearly 50 percentage points higher than Ferrie (1996) while increasing the rate of false positives by only about 10 percentage points.

Our paper is organized as follows. We begin by providing background on the Galveston program. Section III describes the data and our approach to matching individuals over time, including an assessment of our matching algorithm. Section IV illustrates the assignment and eventual location of Galveston immigrants and their demographic and economic characteristics. A simple model of geographic choice is described and implemented in Section V. Section VI briefly concludes.

II. A Brief History of the Galveston Movement

The second half of the 19th and beginning of the 20th century is often referred to as the age of mass migration (Hatton and Williamson 1994; Abramitzky, Boustan, and Eriksson 2012, 2014). Tens of millions of Europeans entered the U.S., with many establishing residency in a

handful of large Northeastern and Midwestern cities. Starting in the 1880s, Jews from Russia began to emigrate in mass as well and likewise generally clustered in those same large cities, most prominently NYC.²

Increasing congestion worried prominent American Jewish philanthropists Jacob Schiff and Baron de Hirsch, who feared the sweeping new waves of Jewish immigrants to NYC would cause nativist backlash and ultimately stricter immigration laws. Moreover, they believed housing and labor market conditions would be better in places that had not experienced the same rapid spike in population. De Hirsch and his philanthropic fund, of which Schiff was a member, initially created the Industrial Removal Office (IRO), tasked to find jobs for Jewish immigrants outside of NYC. The trustees “tried out almost every possible solution - agricultural colonization, suburbanization (on a small scale), the removal of industries to outlying districts, the transportation of families to smaller towns and industrial centers, and so on” (Best 1978, p. 44).

Eventually, Schiff pursued an alternative. Pledging \$500,000 (just under \$13 million in 2016 dollars), Schiff’s plan involved diverting new Russian Jewish immigrants away from NYC.³ The Jewish Territorial Organization (JTO) was charged with recruiting “only young, sturdy immigrants, ready to do whatever work was available” (Best 1978, p. 49) and ensuring those recruits were transported to Bremen, Germany to embark on ships headed to the Port of Galveston. Best (1978) termed the resulting program “Jacob Schiff’s Galveston Movement.”

² See Spitzer (2016) for a discussion of the causes of Russian Jewish emigration.

³ To be clear, the money was not used for trans-Atlantic transportation. Schiff insisted that immigration laws be strictly followed and reprimanded a member of Hilfsverein, a European Jewish aid organization, “for assuring some immigrants that part of their traveling expenses would be paid if necessary. He was willing only that they should be assured that, if the situation required it, the JIIB [Jewish Immigrants’ Information Bureau] would contribute to the expense of their transportation from Galveston to their ultimate destination” (Best, p. 52).

Galveston was chosen for several reasons. First and foremost, Galveston was the farthest port from NYC that serviced Europeans. Logically, destinations farther from NYC were considered more appealing. Moreover, the southeastern states were rejected because of fear of labor market competition from cheap black laborers. Second, Bremen-to-Galveston was already an established shipping route operated by the North German Lloyd line. Finally, Galveston was linked to the railroad network servicing the Midwest and West. The railroads played a key role in Schiff's plan. Upon arrival, immigrants were almost immediately dispersed via the rail lines to pre-determined towns west of the Mississippi River. Destination locations were scouted by the IRO and chosen based on demand for specific jobs (e.g. tailor versus shoemaker) and the size of their Russian Jewish population. Over time, the towns would hopefully build a network of Russian Jews that would make them more appealing to future cohorts.

The first group of immigrants arrived on the *SS Cassel* on July 1, 1907. The 54 program participants were aged 18 to 42 and included “locksmiths, bakers, bookkeepers, noodle and macaroni makers, bookbinders, electricians and shoemakers, as well as many others.” The group was sent to “Colorado, Iowa, Nebraska, Minnesota, Missouri, Illinois, Oklahoma, Texas, and Wisconsin; to sizable cities like Minneapolis, Kansas City, and Milwaukee, and to smaller ones like Davenport, Quincy, and Dubuque” (Best 1978, p. 55).

The Galveston Movement grew from there. Figure 1 provides a time-series of the number of “Hebrew” immigrants that landed in Galveston from 1900 to 1920. Up until 1906, a Hebrew immigrant was rarely routed to Galveston. That changed markedly between 1907 and 1914 (the two vertical lines on the chart), when over 10,000 Jewish immigrants, from more than 7,000 families, arrived. While 10,000 immigrants is a large number relative to the trivial flows prior to 1907, even at its peak, the number of Jewish immigrants arriving in Galveston never

exceeded 4 percent of the annual flow of Russian Jews to the U.S.⁴ The program faced several challenges that restricted its size, including uncooperativeness from the JTO and quality issues with the North German Lloyd ship line. Schiff's program was abruptly terminated in 1914, at which point, he had dispersed roughly half of his original half-million dollar pledge.

III. Data and Matching Methods

A. Data

The Port of Galveston provides a complete list of names, arrival dates, and basic demographic characteristics of the 132,155 passengers that debarked between 1844 and 1949.⁵ We extract 10,076 individuals who departed from the Port of Bremen, Germany,⁶ arrived in Galveston between July 1907 and December 1914, and reported their ethnicity as Hebrew and country of origin as Russia. Our analysis restricts attention to the 5,911 men who were aged 15 to 44 in 1910. We match this sample of Galveston arrivals to the 100 percent population counts from the 1910 to 1940 decennial censuses.⁷ In the next subsection, we detail our matching procedure and assess the quality of matches relative to other reasonable benchmarks. In the meantime, Table 1 provides descriptive statistics about the prime age male Galveston immigrants

⁴ Working-age men were overrepresented among Galveston immigrants. We calculate Galveston represented the port destination of about 7 percent of all Russian Jewish working-age male immigrants during the Movement era. These figures are calculated from a count of the number of Russian Hebrew immigrants arriving in Galveston between 1908 and 1914 divided by the 1920 census count of the number of Russian immigrants who arrived between 1908 and 1914 and report their mother tongue as Hebrew. We think of this as an upper bound since some Russian Jewish immigrants may report their mother tongue as Russian. But it is also possible that Galveston immigrants and other Russian Jewish immigrants may return to Russia at a differential rate.

⁵ See <http://ghf.destinationnext.com/immigration/Search.aspx>. Ethnicity is not included in the version of the data maintained by the Port of Galveston. We collected this information from a version of the data maintained by Ancestry.com.

⁶ Nearly 90 percent of all arrivals to the Port of Galveston during this period departed from Bremen. Most other arrivals departed from Mexican ports but none of these were recorded as Hebrew.

⁷ The 100 percent census files were generously provided to us by the University of Minnesota Population Center via the data collection efforts of ancestry.com. For information on the IPUMS samples, see Steven Ruggles, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek, *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database], Minneapolis: University of Minnesota, 2010.

and, for context, samples of comparable immigrants that match the same demographic profile and arrival time as our Galveston sample (males, aged 15 to 44 in 1910 and arriving between 1908⁸ and 1914). This includes 10,560 non-Russian, non-Hebrew immigrants who likewise departed from the Port of Bremen and immigrated through Galveston (Column 2), and for completeness, all Russian and non-Russian men who immigrated between 1908 and 1914 in the 1920 census (Columns 3 and 4).

Among our main sample of Galveston immigrants, 92 percent declared their U.S. destination as a place west of the Mississippi River, consistent with Schiff's plan.⁹ The most common destination was Texas itself (37 percent) and another 46 percent were assigned to the Midwest, especially Missouri, Iowa, and Minnesota. The remaining 17 percent were spread throughout the Southwest and Pacific regions. Relative to the non-Hebrew Galveston arrivals, the Galveston Hebrew immigrants were slightly younger (25.6 versus 27.6 years), less likely to list their final destination as Texas (37.4 versus 70.3 percent), and appear to be more skilled. Almost half the Hebrew immigrants reported their occupation as craftsman (46.3 percent), whereas 48.4 percent of non-Hebrew arrivals to Galveston were farmers.¹⁰

B. Matching Procedure

Historical record linkage typically relies on phonetic codes, such as the New York State Identification and Intelligence System (NYSIIS), to resolve minor variations in spelling (e.g. Ferrie 1996; Abramitzky et al 2012; Long and Ferrie 2013). Given the ethnic heritage of our

⁸ While the Galveston program began in July 1907, we can only observe year of arrival in the census. Consequently, we only include those arriving between 1908 and 1914.

⁹ While Schiff's wish was to settle families west of the Mississippi River, if sufficient cooperation from a Western location could not be obtained, a family could be sent to states contiguous to the Mississippi River. Of the 8 percent recording destinations east of the Mississippi River, most headed to Illinois (3.6 percent) or Tennessee (1.7 percent).

¹⁰ Other common occupations include operatives (15 percent), farm laborer (8 percent), and clerks (8 percent).

sample, we further address phonetic name coding with the Daitch-Mokotoff (DM) Soundex system (JewishGen). This variant of the Soundex system allows for common variants of Yiddish and Slavic name modifications by matching names according to their pronunciation rather than their spelling.¹¹

Bailey et al (2017) convincingly show that an over-reliance on phonetic codes increases rates of both type I and type II error. Consequently, it is becoming more standard to use phonetic codes solely as a “blocking” mechanism to eliminate non-matches, as in Herzog, Scheuren, and Winkler (2007). That is, we define the universe of potential matches for each Galveston participant as individuals in a census having a DM Soundex match or an NYSIIS match on last name. With our potential set of matches, we then calculate a score for each Galveston-census pair based on Russian origin, age, year of immigration, Russian or Hebrew as native language, and the Jaro-Winkler string similarity scores for first and last names and their interaction.¹² Importantly, the Jaro-Winkler metric allows us to incorporate meaningful variation in name spellings that go beyond simple phonetic matches.

An important decision in matching is how to combine all of this information to select the best possible set of matches. We move beyond the prior literature in calibrating the weight put on our various indicators by exploiting a key feature of the data. In particular, there cannot be matches to the 1910 census for Galveston participants that arrive after 1910.¹³ This observation

¹¹ Our implementation of DM is available upon request. For more information on specific rules, see <http://www.jewishgen.org/InfoFiles/soundex.html#DM>.

¹² Unfortunately, the 1940 census did not ask year of immigration or, universally, native language. Therefore, we expect lower quality matches in 1940 relative to earlier years where those two variables are available. Indeed, as we discuss below, we get higher match rates, as well as higher false-positives, in 1940. Consequently, we put less emphasis on the 1940 match.

¹³ We assume that post-1910 immigrants were not in the U.S. at the time of the 1910 Census (April 1910). We checked for evidence of an earlier migration with the Galveston records that date back to the mid-19th century and were unable to find any examples. That, of course, does not eliminate the possibility of an earlier migration to another port. Our synthetic data tests, described below, are also inconsistent with this concern.

implies that parameters can be chosen to maximize the match rate between the 1907 to 1909 Port records and the 1910 census and minimize the match rate between the 1911 to 1914 Port records and the 1910 census. In addition, the short period from arrival to the 1910 census gives us some further contextual detail that we use to identify potentially false positive matches.¹⁴ We incorporate these insights into the tuning of our matching algorithm, the details of which can be found in Appendix A. Our approach is similar to that of Feigenbaum (2016) who sets parameters in his matching algorithm by training the scoring on a hand-matched subsample of his data. We use this calibrated scoring rule to assign a score to every potential match between a Galveston and a census record.

A second unique feature of our procedure is the use of a deferred acceptance algorithm (DAA) to select the best match for each Galveston record. The DAA matches a Galveston record to the unmatched census record with the highest match score. Consequently, the procedure addresses the possibility that multiple Galveston individuals can simultaneously match to the same census individual and, therefore, by construction create at least one incorrect match. Therefore, the DAA can improve match quality relative to current algorithmic methods that do not address the potential for multiple matches on the same census record.¹⁵ Appendix A describes the DAA algorithm in more detail. Matches with non-positive scores or extreme differences in ages are discarded.¹⁶

¹⁴ This additional detail involves flagging matches if a) he reports citizenship in the 1910 census (not possible for 1907-09 immigrants), b) he report no children in the 1910 census if the individual arrived in Galveston with children, or c) the potential match results in a difference of more than 5 years in age or arrival in the U.S.

¹⁵ To be concrete within our context, we believe the current literature focuses on multiple census records being linked to the same Galveston record and ignores multiple Galveston records being linked to the same census record. To our knowledge, the only matching paper to consider this feature in the specification of their algorithm is Feigenbaum (2016).

¹⁶ The cutoff for differences in ages is a calibrated parameter. We fix the cutoff for overall score; the inclusion of a constant parameter in the creation of the score means this is in effect a flexible cutoff. Intuitively, we expect a higher constant to reflect general uncertainty in the correct match conditional on being a phonetic match.

C. Alternative Procedure

We assess the performance of our preferred match algorithm relative to the widely-used procedure described in Ferrie (1996) and Long and Ferrie (2013).¹⁷ This method has the advantage of minimizing false positives but at the cost of a smaller and less representative sample. We implement the Ferrie matching method as follows:

1. Consider anyone who has an NYSIIS match on last name in the census.
2. Restrict to men older than 10 at arrival in the Galveston data and older than 10 in the census.
3. Drop if the difference in ages is greater than 5.
4. Drop if the difference in arrival year is greater than 5.
5. For 1910 and 1920 only, drop if the individual has children in the Galveston records but does not have children in the census.
6. Drop any Galveston individual who has more than 10 potential census matches at this point.
7. Take the individual for whom there is the smallest difference in ages and then the smallest difference in year of arrival. If there are multiple candidates for this match, drop the Galveston individual and any potential census matches.

We depart slightly from the restrictions imposed by Ferrie (1996) in order to adapt his algorithm to our immigrant setting. These differences are: a) excluding restrictions on state of birth, b) swapping household head status with presence of children (step 5), c) setting the maximum age differences to 5, which lies between the restrictions in Ferrie (1996) and Long and Ferrie (2013), and d) moving step 6 to after steps 3 to 5. Dropping those with many potential matches and with

¹⁷ We discuss how our results vary by matching procedure in section V.B.

duplicate final matches (steps 6 and 7) decreases sample sizes, often considerably, and may decrease representativeness of the final sample since common names are those that typically get removed.

D. Assessing the Matches

We assess the quality of our matches in two ways.

First, we compare the match rates to the 1910 Census for those arriving prior to 1910 with those arriving after 1910. Using the full set of data, the pre-1910 arrival match rate is 73 percent and the post-1910 rate is 11 percent. These match rates imply the probability of a false-positive *conditional on matching*, i.e. the type I error rate, is $11.2/73.4 = 15.2$ percent. As an alternative exercise, we calibrate our scoring procedure using half our data and evaluate the matches using a hold-out sample. In that case, the pre-1910 match rate in the hold-out sample is 76.0 percent and the post-1910 match rate is 13.4 percent. Together, these imply a type I error rate of 17.7 percent. By comparison, Bailey et al (2017) estimate type I error rates of between 18 and 70 percent for some of the prominent algorithmic methods in the literature. In our main analysis, we drop these false-positive post-1910 matches.

Our second test matches the Galveston data to a constructed synthetic dataset consisting of randomly perturbed Galveston records combined with known non-matches, a method used in Bailey et al (2017).¹⁸ Since Galveston data are included in both the master and synthetic datasets, we know if the final match is “true,” allowing us to assess the rate of type I errors. See Appendix B for details on how this exercise is constructed.

¹⁸ The known non-matches consist of Russians arriving between 1900 and 1905 and between 1916 and 1920 with arrival year imputed for added robustness. Our perturbation to the Galveston data closely follows Bailey et al (2017) with a few amendments (see Appendix B).

The 1910 through 1930 match rates for both the final data used in our analysis (Column 1) and the synthetic data (Column 2) are shown in Table 2. Match rates are, unsurprisingly, significantly higher for our preferred method compared to the benchmark Ferrie method. The synthetic test implies that up to 35.3 percent of our preferred matches to the 1920 census could be false.¹⁹ However, we believe that this number likely represents an upper limit on the false positive rate due to the great deal of noise, such as randomly assigning arrival years to the non-matched Russian sample, introduced into the records. Indeed, the synthetic test's error rate is at least twice that of the implied 15 to 18 percent false positive rate based on the 1910 census match tests described earlier in this section. Moreover, we have no way to assess whether the level of measurement error introduced to the synthetic data reflect similar noise in the real census data.

Therefore, of more interest is the performance of our algorithm *relative* to existing methods rather than the absolute error and match rate. Taken together, the synthetic tests in Table 2 imply that our method drastically improves match rates relative to the Ferrie method – 70 to 73 percent compared to 23 to 25 percent -- while somewhat increasing the rate of false-positive matches from about 23 percent using the Ferrie method to 35 percent with our preferred method.

Our preferred method uses the single best match for each Galveston participant. We also tried incorporating match uncertainty by including up to five census matches with the highest scores for each person. Each match is weighted by the size of the match score, with the weights for each participant summing to unity. In Table 2, we call this method the Five Best Weighted

¹⁹ We compute a somewhat smaller match rate but a very similar false positive rate if we drop the small share of Galveston residents that initially were assigned east of the Mississippi River.

Matches. We find that the Five Best match rates are similar to our preferred method but with a substantially higher rate of false-positives.

IV. Location and Characteristics of Matches

We start with an overview of where the Galveston participants were initially assigned when they arrived in the U.S. and then show where they are located as of the 1920 census.

Figure 2 displays the assignment location of Galveston immigrants both in the full sample (Panel a) and conditional on being matched to the 1920 census using our preferred match procedure (Panel b). The states of Texas, Iowa, Minnesota, and Missouri have the largest shares at respective rates of 37.4, 11.8, 10.9, and 10.1 percent. That these patterns do not change when we condition on matching (comparing Panels a to b) shows that there is little selection on assignment location.

Figure 3 shows the 1920 locations of Galveston immigrants matched to the 1920 census using the Preferred, Five Best, and Ferrie methods, respectively (Panels a to c), and of a comparison sample of Russian-born men who also arrived in the U.S. between 1908 and 1914 drawn from the 1920 census (Panel d).²⁰ The maps highlight three striking features. First, rather than staying west of the Mississippi River, Galveston immigrants largely moved east, and especially to the Northeast and Illinois (we quantify this observation below). Second, the locations chosen are highly correlated with those of the Russian-born/non-Galveston men plotted

²⁰ Appendix Figures A1, A2, and A3 show analogous maps for 1910, 1930, and 1940.

in Figure 3d. Indeed, the state-level correlation of the locations of Galveston immigrants and the Russian men is 0.97.²¹ Third, these patterns are robust to matching criteria.

While Galveston participants gravitate to states east of the Mississippi River with high Russian immigrant populations, this does not mean that assignment to a state lacks empirical content. Figure 4 combines the data embedded in Figures 2 and 3 to show the percentage of those remaining in their initially assigned state as of the 1920 census.²² While assignment compliance is overall low, it also varies substantially across states, and again this observation is irrespective of matching method. For example, California has a compliance rate of 12.4 percent using our preferred method while Texas's rate is 2.4 percent.

Table 3 complements the maps by showing summary statistics for our preferred matches at the time of their arrival in Galveston (Column 1) and in the 1910 to 1940 censuses (Columns 2 to 5). Validating our sample and match procedure, we find a) age increases by roughly 10 years between each decennial census²³, b) the percent of the sample arriving prior to 1910 remains constant at 18 percent and corresponds to the unconditional rate at arrival displayed in Column 1, and c) certain variables such as marriage, naturalization, literacy in English, home ownership, self-employment, and managerial occupation display monotonic trends consistent with aging and assimilation with longer tenures since arrival.²⁴ Interestingly, higher skilled jobs are much more common among the Galveston immigrants than the cohort of non-Galveston Russian immigrant that arrived during the same period (see Table 1).

²¹ See Appendix Figure A4. The correlation remains that high even if we discard New York.

²² These figures ignore the small number of immigrants initially assigned east of the Mississippi River as well as states receiving five or fewer arrivals.

²³ This pattern does not hold for the 1910 matches, since the sample is partly selected on pre-1910 arrivals, who tend to be older than their post-1910 counterparts.

²⁴ For matches to the 1910 census, rather than 100 percent, 54.1 percent arrive prior to 1910 since this calculation is not inclusive of the year 1910 itself.

Table 3 also quantifies where Galveston’s immigrants ultimately were at each decennial census through 1940. Of the 88.4 percent that moved east by 1920,²⁵ over 40 percent (37.9/88.4 percent) lived in New York and about 60 percent (52.4/88.4 percent) in the NY-NJ-PA region. Those NY rates exceed that of Russian-born/non-Galveston men who arrived between 1908 and 1914, a discrepancy possibly explained by our inability to calculate them specifically by Hebrew status among the non-Galveston Russian men. While only 11.6 percent live west of the Mississippi River in 1920, this number steadily increases over time to 18.5 percent by 1940, a trend driven by movement to California. These descriptive results mirror trends highlighted in Bartel (1989), who finds that post-1964 immigrants initially move into areas with high concentrations of their own ethnicity but become less sensitive to ethnic enclaves over time.

V. Explaining Mobility Choices

A. Conceptual Framework and Empirical Strategy

To understand how demographic and economic characteristics influence residential mobility choices, we use a model pioneered by McFadden (1974) and previously applied to migration decisions by Dahl (2002). We assume individual i ’s net utility from living in location k is additively separable in the individual’s earnings y in location k , cost of moving c to location k , and tastes t over location k :

$$(1) \quad U_{ik} = y_{ik} - c_{ik} + t_{ik}$$

²⁵ For clarity, we define Louisiana and Minnesota as west of the Mississippi River, even though the river partially cuts through these states.

where t_{ik} encompasses non-wage location-specific factors such as demographic makeup, public amenities, geography, and weather. We can rewrite equation (1) as the sum of the average utility from moving to location k , V_k , and an individual specific utility, v_{ik} :

$$(2) \quad U_{ik} = V_k + v_{ik}.$$

where V_k is a hedonic function of the location's characteristics:

$$(3) \quad V_k = E[y_{ik} - c_{ik} + t_{ik} | x_i, z_k, Assigned_{ik}] = \beta(x_i) \cdot z_k.$$

x_i is a vector of individual characteristics and z_k is a vector of location characteristics. The individual specific component, v_{ik} , can be decomposed as follows:

$$v_{ik} = \delta Assigned_{ik} + \varepsilon_{ik},$$

where $Assigned_{ik}$ indicates whether individual i is assigned to location k . δ can be interpreted as an individual's net value of assignment to state k . ε_{ik} represents all other idiosyncratic factors affecting individual i 's latent utility over state k .

Individuals select whichever location yields the greatest utility. Let M_{ik} be an indicator for individual i choosing location k . Specifically, M_{ik} equals 1 if:

$$(4) \quad V_k + \varepsilon_{ik} \geq V_{k'} + \varepsilon_{ik'} \text{ for all } k' \neq k.$$

Otherwise, M_{ik} equals 0.

This simple model implies that a migration program like Galveston will shift an individual's location choice from k to k' when:

$$(5) \quad \delta Assigned_{ik'} \geq \beta(x_i) \cdot (z_k - z'_k) + \varepsilon_{ik} - \varepsilon_{ik'}.$$

Equation (5) has three implications about the importance of initial assignment on an individual's ultimate location choice. First, location decisions are likely altered when δ is large. Second, when an individual is assigned to a location k that differs from k' along feature j , individuals will only change their location if j is relatively unimportant, or specifically $\beta^{(j)}(z_k^{(j)} - z_{k'}^{(j)})$ is small. Finally, assignment is more likely to affect behavior when an individual's preference for a particular location is less idiosyncratic, or $\varepsilon_{ik} - \varepsilon_{ik'}$ is small.

Our aim is to estimate Galveston participants' average latent utility to choosing any of the 48 U.S. states.²⁶ We observe a vector of individual and state characteristics, x_i and z_k respectively, and where each individual chooses to locate, M_{ik} . Under the strong assumption that all relevant state characteristics z_k are observed, the model implies:

$$(6) \quad \beta(x_i) \cdot (z_k - z_{k'}) + \delta(\text{Assigned}_{ik} - \text{Assigned}_{ik'}) \geq \varepsilon_{ik} - \varepsilon_{ik'} \quad \forall k'$$

Point identification of the coefficients $\beta(x_i)$ and δ is challenging without parametric assumptions about the distribution of the idiosyncratic component of utility, ε_{ik} . Therefore, like the much of the literature since McFadden (1974), we assume ε_{ik} are independent and identically distributed according to the Type I Extreme Value Distribution. With this assumption, we can express the probability individual i chooses state k as:

$$(7) \quad P(M_{ik} = 1 | x_i, z_i, \text{Assigned}_{ik}) = \frac{\exp \beta(x_i) \cdot z_k + \delta \text{Assigned}_{ik}}{\sum_{k'} \beta(x_i) \cdot z_{k'} + \delta \text{Assigned}_{ik'}}$$

Equation (7) is McFadden's alternative-specific conditional logit model which allows covariates to vary by individual and state jointly.

²⁶ Of course, Alaska and Hawaii were not states at this time. Since we do not have some key data for DC, we exclude it as a destination location. Practically, when we include DC, it has no empirical impact as only 17 Galveston participants migrated there by 1920.

B. Results

Table 4 displays estimates of Equation (7) where the outcome is a vector of indicator variables for residing in state k in the 1920 census. The parameters reported in the table represent the impact of a variable on the expected latent utility of living in a particular state. Given our parametric assumptions, this latent utility is interpretable as the log-odds of residing in a particular state in 1920. Point estimates are in the top row and standard errors are in parentheses. For interpretability, we stress the average marginal effects of changing a variable on the propensity to choose that state, which are shown in brackets.²⁷ We also display likelihood ratio tests for the following groups of covariates: (1) total population demographics (labeled “Population”), (2) demographics specific to cohorts (“Cohort”), (3) occupation, and (4) other Galveston immigrants.

The first row quantifies the main treatment effect of the Galveston program: the extent to which assignment to a particular state increases the odds of staying in that state. Controlling for state-specific intercepts only (Column 1), we estimate that the effect of being assigned to a state increases the log odds of selecting a state by a factor of 1.101. This point estimate fluctuates narrowly between 0.724 and 1.233 using different specifications (Columns 2 onwards) and alternative matching procedures (Appendix Table A2).²⁸ This is a big effect. The point estimate

²⁷ Marginal effects are calculated as a matrix of cross-alternative parameters. Specifically, marginal effects for a single covariate are represented as a 48-by-48 symmetric matrix, where each i,j cell is the effect of a change in a covariate in state i on the probability of choosing state j . For interpretability, we take the average of the diagonal in the calculated marginal effects matrix. This makes the marginal effects in brackets interpretable as the average effect of changing a parameter on the probability of choosing a typical state.

²⁸ In addition, we find that our results are robust to restricting the potential census match universe and other minor alterations to the baseline matching procedure. Out of concern that our sample reflects the locations of Russians in general rather than the location of true matches, we exclude potential matches to New York. We also extend this exercise by excluding New York, New Jersey, and Pennsylvania, as well as dropping Galveston arrivals who were assigned east of the Mississippi river. Second, out of concern that the DAA method induces low-quality matches on the second pass, we stop the matching algorithm on its first iteration (so each census record is uniquely matched to

in our most saturated model in Column (11), 1.086, implies that being assigned to a particular state increased the latent utility of that state by the same amount as increasing the number of Russian immigrants, which we will show is the strongest driver of this latent utility, by a factor of 20.²⁹ However, this effect was only a small victory for Schiff, as the vast majority of Galveston families still moved east. The corresponding average marginal effect of assignment on the probability of living in the assigned state in 1920 fluctuates narrowly between 1.3 and 2.2 percent.

How did they choose where to go? Conditional on assigned state, we find that Galveston participants flow to states with larger Jewish and Russian-born immigrant populations.³⁰ Using a sparse specification in column 3, we find that a 10 percent increase in the Russian immigrant or Jewish population increases the propensity to move to that state by 10.7 and 3.1 percent, respectively, and both of these effects are highly statistically significant. Even in our most saturated model shown in column (11), which conditions on a host of other demographic and occupational covariates and the locational choices of other Galveston participants, we still find the same 10 percent increase in the Russian immigrant or Jewish population increases the propensity to move to that state by 6.2 and 1.2 percent, respectively.

However, these latter estimates likely understate the degree to which location decisions are associated with ethnic enclaves since we cannot directly measure the most pertinent population – Russian Jews. Indeed, much of the decline in the Russian and Jewish point estimates between the sparse column 3 and the more saturated specification in column 11 occurs

the Galveston record with its highest score) as well as take the highest score and drop duplicate census matches. None of these exercises have an impact on the results.

²⁹ In particular, $1.086 = 0.361 \log\left(\frac{B}{A}\right)$ which implies $\frac{B}{A} = 20.1$.

³⁰ State estimates of religion are taken from the Census of Religious Bodies. See Department of Commerce and Labor (1910).

when we add a reasonable proxy for Russian Jewish descent -- the location of Russian Jews that immigrated on the Galveston ships. We define $\log(\text{Num. Non-Fam, Located } 1920)_k$ as the log level of *non-family* Hebrew arrivals to Galveston between 1907 and 1914 that live in a particular state k at the time of the 1920 census. It is important to stress that these results are not interpretable as peer effects in the causal sense but simply quantify the degree to which Galveston individuals (and perhaps Russian Jews arriving in the Western US at the beginning of the 20th century) move to similar destinations (Manski 1993).³¹ That said, we find a 10 percent increase in the population of Galveston immigrants in 1920 is associated with an 11.1 percent increase in the propensity to move to that state.³² Controlling for the number of other Galveston immigrants in state k also eliminates the effect of the total immigrant population on location choice, further indicating that location decisions are driven by well-defined ethnic enclaves. These results are robust to other matching methods (Appendix Table A2).³³

We find mixed evidence that location decisions were associated with the human capital of locals, as proxied by the separate male and female 1910 literacy rates for both native-born and immigrant populations from approximately the same ten year birth cohorts (Columns 6, 10, and 11). There is some indication that Galveston participants were more likely to move to states with higher literacy rates of male immigrants although this result disappears when we control for non-family Hebrew arrivals from the Galveston program. In our more saturated specifications

³¹ Our use of the multinomial logit model induces a non-linearity into the agent's decision process that allows the Galveston immigrant effects to be estimated (Brock and Durlauf 2002). We include a dummy indicating that a state has no non-family Galveston immigrants. The results are similar whether we look at immigrants arriving in the same year or over the full 1907 to 1914 period.

³² We also tried using the 1920 location of shipmates, defined by the variable *Located 1920, Ship (%)* as the percent of *non-family* passengers on the same ship as participant i that moved to a particular state k . While the location of shipmates is somewhat associated with location choice (Column 4), the effect is economically small and insignificant when we condition on other covariates (Column 11).

³³ We also looked at 1930 location decisions. Aside from some expected attenuation in the effect of assignment, the results are nearly identical for our more complete specifications.

(Columns 10 and 11), we find that Galveston immigrants were significantly less likely to reside in areas with native residents with higher literacy rates. There is also some, but not especially robust, evidence that prime age male Galveston participants are more likely to move to states with more similarly aged women.

While we find that locations were driven by clusters of ethnic enclaves, we do not want to dismiss the role of economic opportunity. However, the importance of income and occupation are mixed and of secondary economic importance to ethnicity. Adding the log of the state's mean occupational income³⁴ to our statistical model, we find that states with better opportunities detract (Column 7) or at least have no impact (Column 11) on the location choices of Galveston immigrants. Similarly, there is no evidence that migration decisions are affected by the share of jobs in their specific two-digit occupation at arrival (Columns 8 and 11). The best evidence that we have uncovered that economic opportunities affected decisions is that Galveston men tend to move to states with a higher percentage of higher-paid professional workers and craftsmen and away from states with a higher fraction of lower-paid farmers and clerical/sales/service workers (Columns 9 to 11).³⁵ After controlling for other factors, a 10 percent increase in the share of professionals and craftsmen is associated with a 1.0 and 2.4 percent increase, on average, in the probability of moving to a state.

Finally, we split the sample into thirds based on occupational earnings scores at the time of arrival in Galveston (Table 5). State assignments increased the probability of living in a state by about 2 percentage points across all three groups. Similarly, there are small, typically statistically insignificant differences in the response to the number Russian immigrants and

³⁴ Occupational income is IPUMS' *OCCSCORE* variable, which calculates median earnings by occupation in 1950.

³⁵ These shares roughly correspond to 1-digit occupation codes. We exclude shares for general laborers and non-occupational responses in the estimates.

Galveston peers across the three occupational earnings terciles. The main difference between the groups is that the occupation effects are economically more important among the higher skilled workers, suggesting some heterogeneity in the relevance of economic opportunity. In particular, individuals in the top tercile are much more attracted to states with a large share of professionals and craftsmen, even after controlling for the location of other Galveston arrivals. Nevertheless, the appeal of being in an ethnic enclave is present and dominant across the skill distribution.

VI. Conclusion

This paper studies an unusual natural experiment that assigned initial U.S. destination to roughly 10,000 immigrants just prior to World War I. To alleviate congestion in New York City, philanthropists recruited Russian Jewish migrants to board ships in Bremen, German destined for Galveston, Texas. Upon arrival, families were given train tickets to locations west of the Mississippi River where a job awaited. Since participants had no influence on their initial assignment, the Galveston Movement had many features of later mobility programs such as Moving to Opportunity. Yet despite the stated purpose of the program to locate Russian Jewish immigrants to ethnically sparse areas of the U.S., we find that almost 90 percent of the prime age male participants in this program ultimately moved east of the Mississippi, and typically to large Northeastern and Midwestern cities.

To explain this lack of compliance, we use a multinomial logit framework inspired by a Roy model for location decisions. We show that Galveston immigrants were ultimately attracted to enclaves with similar religious and country of origin background. By contrast, there is no economically or statistically significant effect from a place having a larger base of immigrants from other areas of the world. Moreover, while we find an association between observable

economic conditions and location, these appear to be of secondary importance to our ethnic measures, especially for participants in the bottom part of the skill distribution. We interpret this evidence as consistent with ethnic and religious networks playing a vital role in location decisions for early 20th century Russian Jewish immigrants, paralleling the experiences of many other immigrant waves to the U.S. over the remainder of the 20th century and perhaps to the experiences of the Moving to Opportunity program much later (Orr et al. 2003).

Finally, our paper introduces two novel adjustments for matching census data. First, we tune our scoring of potential matches based on an objective measure of match quality: the share of Galveston immigrants who arrived after 1910 who were matched to the 1910 census. Second, we use a deferred acceptance algorithm (DAA) to avoid the real possibility that multiple Galveston immigrants are matched to the same individual in a Census. Using the diagnostics developed in Bailey et al (2017), we show that our approach performs favorably relative to standard algorithmic approaches popular in the literature.

References

- Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson, 2014, "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 122(3): 467-506.
- Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson, 2012, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 102(5): 1832-1856.
- Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, 2017, "How Well do Automated Linking Methods Perform in Historical Samples? Evidence from New Ground Truth," NBER Working Paper no. 24019.
- Bartel Anne, 1989, "Where Do the New U.S. Immigrants Live?" *Journal of Labor Economics*, 7(4): 371-391.
- Beaman, Lori, 2012, "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettles in the U.S.," *Review of Economic Studies* 79: 128-161.
- Best, Gary, 1978, "Jacob H. Schiff's Galveston Movement: An Experiment in Immigrant Deflection, 1907-1914," *American Jewish Archives*, April: 43-79.
- Brock, William, and Steven Durlauf, 2002, "A Multinomial Choice Model with Neighborhood Interactions," *Review of Economic Studies*, 68: 235-260.
- Cutler, David M., Edward Glaeser, Jacob Vigdor, 2008, "When are Ghettos Bad? Lessons from Immigrant Segregation in the United States," *Journal of Urban Economics*, 63: 759-774.
- Dahl, Gordon, 2002, "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70(6): 2367-2420.
- Damm, Anna Piil, 2009, "Ethnic Enclaves and Immigration Labor Market Outcomes: Quasi-Experimental Evidence," *Journal of Labor Economics*, 27(2): 281-314.
- Damm, Anna Piil and Christian Dustmann, 2014, "Does Growing Up in a High Crime Neighborhood Affect Youth Criminal Behavior," *American Economic Review*, 104(6): 1806-1832.
- Department of Commerce and Labor, Bureau of the Census, 1910, *Census of Religious Bodies, County File, 1906*, Washington, D.C.: Government Printing Office, retrieved from: <http://www.thearda.com/Archive/Files/Descriptions/1906CENSCT.asp>
- Edin, Per-Anders, Peter Fredriksson, and Olof Aslund, 2003, "Ethnic Enclaves and the Economic Success of Immigrants – Evidence from a Natural Experiment," *Quarterly Journal of Economics* 118(1), 329-357.
- Feigenbaum, James, 2016, "A Machine Learning Approach to Census Record Linking," Working Paper, Harvard University.

- Ferrie, Joseph, 1996, "A New Sample of Males Linked from the Public Use Micro Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 29(4): 141-156.
- Gale, David, and Lloyd Shapley, 1962, "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69(1): 9-15.
- Haines, David, , 1989, *Refugees as Immigrants: Cambodians, Laotians, and Vietnamese in America*, Totowa, N.J.: Rowman and Littlefield.
- Hatton, Timothy, and Jeffrey Williamson, 1994, "What drove the mass migration from Europe in the Late Nineteenth Century?" *Population and Development Review*, 20(3): 533-559.
- Herzog, Thomas N., Fritz Scheuren, and William Winkler, 2007, *Data Quality and Record Linkage Techniques*.
- JewishGen. "Soundex Coding". Accessed July 16th, 2014.
<http://www.jewishgen.org/infofiles/soundex.html>
- Long, Jason, and Joseph Ferrie, 2013, "Intergenerational Occupational Mobility in Great Britain and the United States since 1850," *American Economic Review*, 103(4): 1109-1137.
- McFadden, Daniel, 1974, "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, ed. P. Zarembka, 105-142.
- Manski, Charles, 1993, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60: 531-542.
- Nix, Emily and Nancy Qian, 2015. "The Fluidity of Race 'Passing' in the United States, 1880-1940," NBER Working Paper no. 20828.
- Orr, Larry, Judith Feins, Robin Jacob, Erick Beecroft, Lisa Sanbonmatsu, Lawrence Katz, Jeffrey Liebman, and Jeffrey Kling, 2003, *Moving to Opportunity for Fair Housing Demonstration Program: Interim Impacts Evaluation*, US Department of Housing and Urban Development, Office of Policy Development and Research.
- Robinson, W. Courtland, 1998, *Terms of Refuge: The Indochinese Exodus and the International Response*, London: Zed Books.
- Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew Schroeder, and Matthew Sobek, 2010, *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database], Minneapolis: University of Minnesota.
- Spitzer, Yannay, 2016, "Pogroms, Networks, and Migration: The Jewish Migration from the Russian Empire to the United States, 1881-1914," Working Paper, Hebrew University.

Appendix A: Additional Detail on the Matching Algorithm and Calibration

Once we obtain the universe of potential matches, defined as having either a DM Soundex or NYSIIS phonetic match, we construct a match score for each record pair. Let g denote a Galveston record and c denote a census record. The match score is:

$$\begin{aligned}
 SCORE_{cg} = & \beta_{Constant} + \beta_{Born\ Russian} D[Born\ Russian]_c + \beta_{Age} |Age_c - Age_g|^{\gamma_{Age}} \\
 & + \beta_{Year\ Arrive} |Year\ Arrive_c - Year\ Arrive_g|^{\gamma_{Year\ Arrive}} \\
 & + \beta_{Last\ JW_{Last}} + \beta_{First\ JW_{First}} + \beta_{Last\ X\ First\ JW_{Last}\ JW_{First}} \\
 & + \beta_{Speak\ Russian} D[Speak\ Russian]_c + \beta_{Speak\ Hebrew} D[Speak\ Hebrew]_c
 \end{aligned}$$

where $D[-]_c$ indicates a dummy for having a characteristic in the census data and JW_- indicates Jaro-Winkler string similarity score for names. $\gamma_{Year\ Arrive}$ and γ_{Age} are parameters that penalize difference in age or year of arrival by changing the convexity of the difference.

We tune the parameters used in the algorithm by optimizing an objective function that maximizes the pre-1910 match rate and minimizes the post-1910 match rate. To avoid overfitting, for example on year of arrival, we also minimize the rate of “bad matching” based on matches having implausible characteristics in the 1910 census. Specifically, this is defined as being a U.S. citizen, having children in the Galveston data but not the census, and having differences in age or year of arrival of more than 5 years. The specific objective function we maximize is:

$$f(e, l, b) = e(e - l^{1.5})(1 - b^{1.5})$$

This is optimized by random search over the fourteen parameters the algorithm depends on. Specifically, we run the algorithm for 250 iterations, with each iteration consisting of 200 random draws. The first 60 random draws sample values for all parameters jointly. The latter 140 draws iteratively sample 10 draws for each for the parameters. Each parameter is given a bandwidth to search in; if none of the draws in an iteration of the algorithm improve the performance of the objective function, the bandwidth for the parameters is decreased by a factor of 0.99.

We calibrate the algorithm over the sample of all white foreign-born males who are aged 10 or older upon arrival in the Galveston data and 10 or older in the census data. While we calibrate these parameters for all white immigrants, our baseline matches are restricted to Russian-born, white immigrants aged 25 to 54 in 1920.

We run the optimization for several starting values and random seeds. Unfortunately, the non-concave nature of the matching function means that convergence is not unique. We choose the final values based on the parameter set which corresponds best to the optimization where we only use Russian-born individuals from the census. Table A1 shows the final values and upper and lower limits of the search area for each parameter used to create our matches.

Once match scores are created, we drop non-positive scores as well as records where the absolute value of the difference in ages is more than δ_{Age} , an additional parameter for which we solve. While dropping pairs with a non-positive score is somewhat arbitrary, the inclusion of a constant, $\beta_{Constant}$, means that the cutoff is in reality flexible.

Next, we implement a deferred acceptance algorithm to find the final match pair using the match score as the numerical strength of the potential matching. The algorithm iterates until

all Galveston individuals have a match or no longer have a potential match. The procedure is as follows:

1. On iteration S , match each Galveston record to its highest score, $SCORE_{cg}$.
2. If a census record is matched to multiple Galveston records, take the Galveston record with the highest $SCORE_{cg}$. Score ties are broken by choosing the record pair with the smallest age difference and the Galveston record for which the census record represents the largest increase in score relative to the next best census record. Delink the Galveston records with lower scores for this census record.
3. Save these unique matches and remove any Galveston or census record that is included in this set.
4. Repeat the algorithm with the remaining Galveston-census combinations that have thus far not been included in the matches.

These rules will create a set of unique matches representing the best combination for all records.

Appendix B: Synthetic Validation

Our synthetic validation procedure follows Bailey et al (2017). We create a synthetic dataset of perturbed true matches and known non-matches which we then match to the original, unperturbed data. This allows us to compare match rates as well as error rates in both the original and synthetic dataset. This test is performed over the sample of males aged 10 and older.

Creating the synthetic dataset involves two steps. First, for a subset of the data, we randomly introduce measurement error to the Galveston data consistent with Bailey et al (2017).

This takes a number of forms:

1. **Names:** Randomly switching first and last names and introducing duplication, deletion, and transposition of characters in the first and last name.
2. **Age:** Randomly round 25 percent of ages to the closest multiple of 5 to simulate age clumping.
3. **Year of arrival:** Randomly change 10 percent of arrival years to within 5 years of the original year listed, randomly shift 5 percent to the previous decade, and randomly “clump” 5 percent by rounding to the nearest year multiple of 5.
4. **Mother tongue:** Mother tongue is not a variable in the Galveston data. We randomly assign Russian, Hebrew, or neither as mother tongue based on the empirical proportion of Russian-born migrants from 1907-1914 reporting these languages (conditional on reporting Russian, Hebrew, English, or no response). To this number, we randomly change 20 percent to neither Russian nor Hebrew.

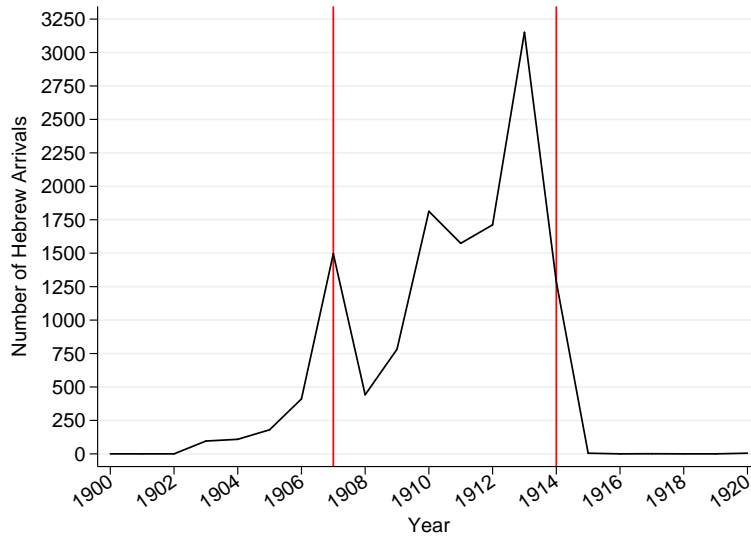
5. **Non-matches:** Following Bailey et al (2017), individuals who die, move out of the U.S., or are unable to be matched for other reasons are simulated by dropping 15 percent of the true matches.

In the second step, we construct a set of census individuals who should be non-matches. This group needs to be both comparable enough that they have the potential to be falsely matched and different enough that we know *a priori* that they should not be matched. Consequently, we select Russian-born men who arrive in Galveston between 1900 to 1905 or 1916 to 1920. Since our algorithm depends heavily on matching year of arrival, the synthetic test shows very low error rates when we do not alter the non-match sample's year of arrival. To create additional difficulty for our algorithm, we randomly assign a year of arrival between 1907 and 1914 for all known non-matches. This addition drastically increases error rates in the synthetic test.³⁶

The two sets of perturbed true matches and known non-matches are appended together to create synthetic census data. We then match the original Galveston data to this synthetic census dataset using the methods described already. These results are displayed in Table 2 and discussed in the main text.

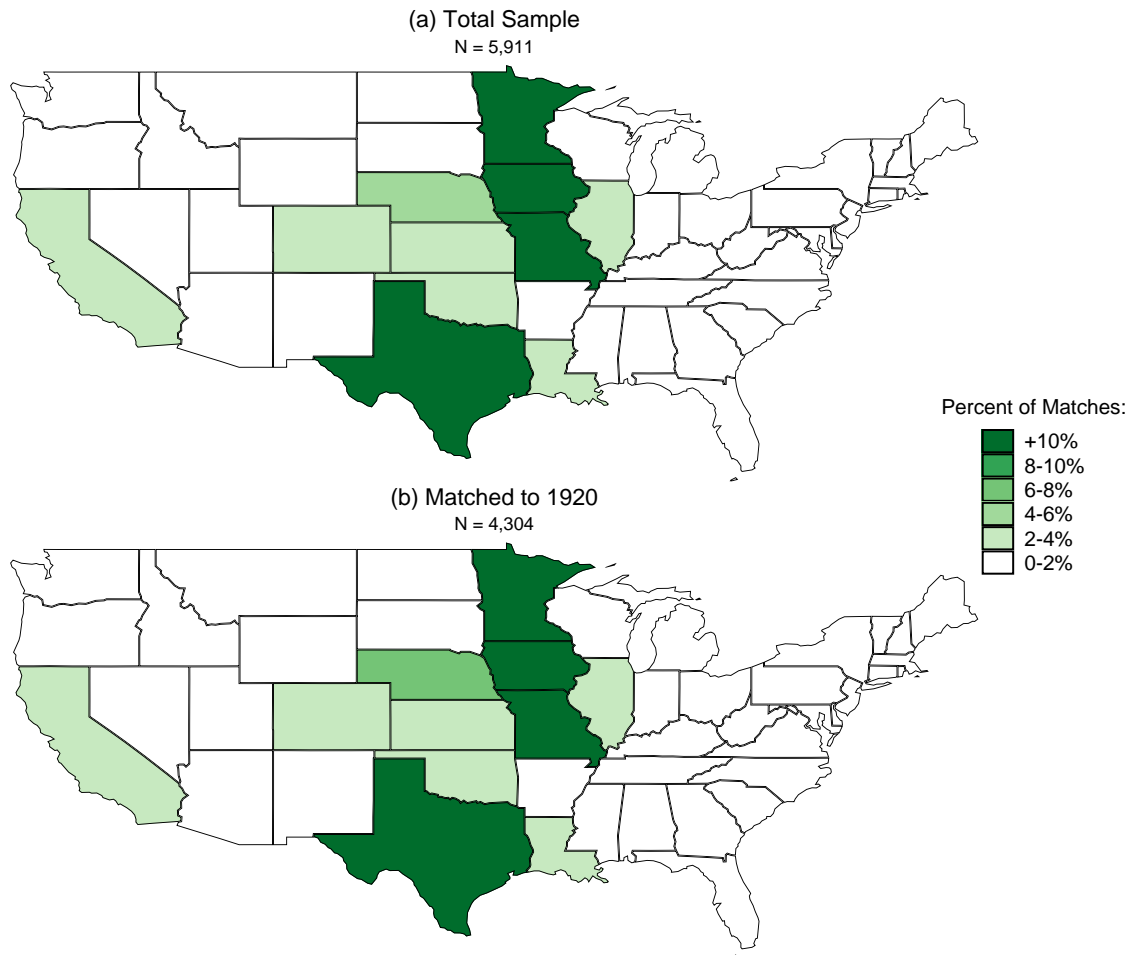
³⁶ Without allocating year of arrival, all error rates are below 5 percent for our preferred algorithm.

Figure 1: Number of Hebrew Arrivals to the Port of Galveston: 1900-1920



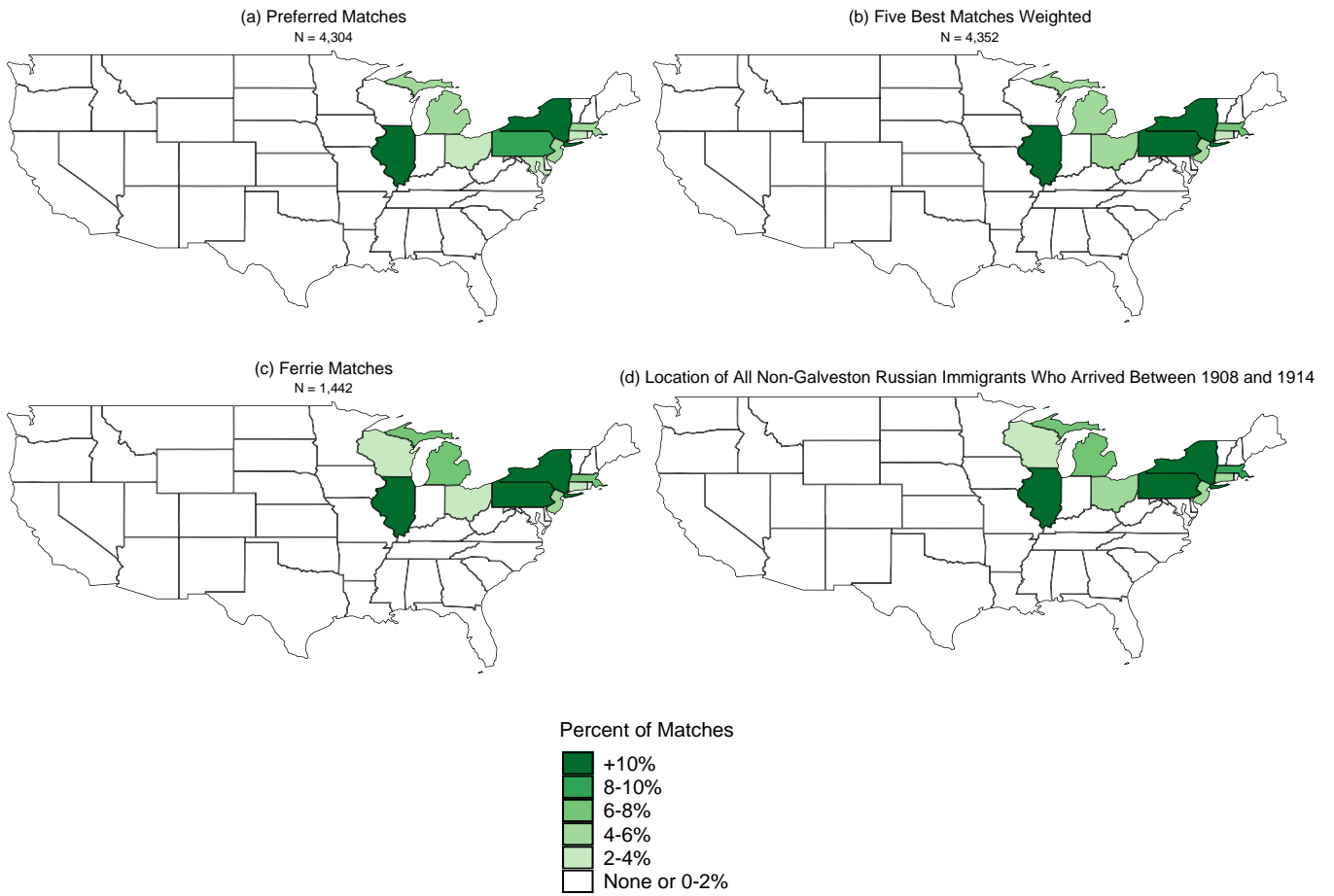
Notes: This figure plots the number of arrivals tagged as “Hebrew” (Russian and non-Russian) arriving to the port of Galveston between 1900 and 1920. The red lines are at 1907 and 1914, indicating the start and end of the Galveston program, respectively. Source: Ancestry.com.

Figure 2: Assignment State After Galveston Arrival, Men Born Between 1866 and 1895 (Aged 15 to 44 in 1910)



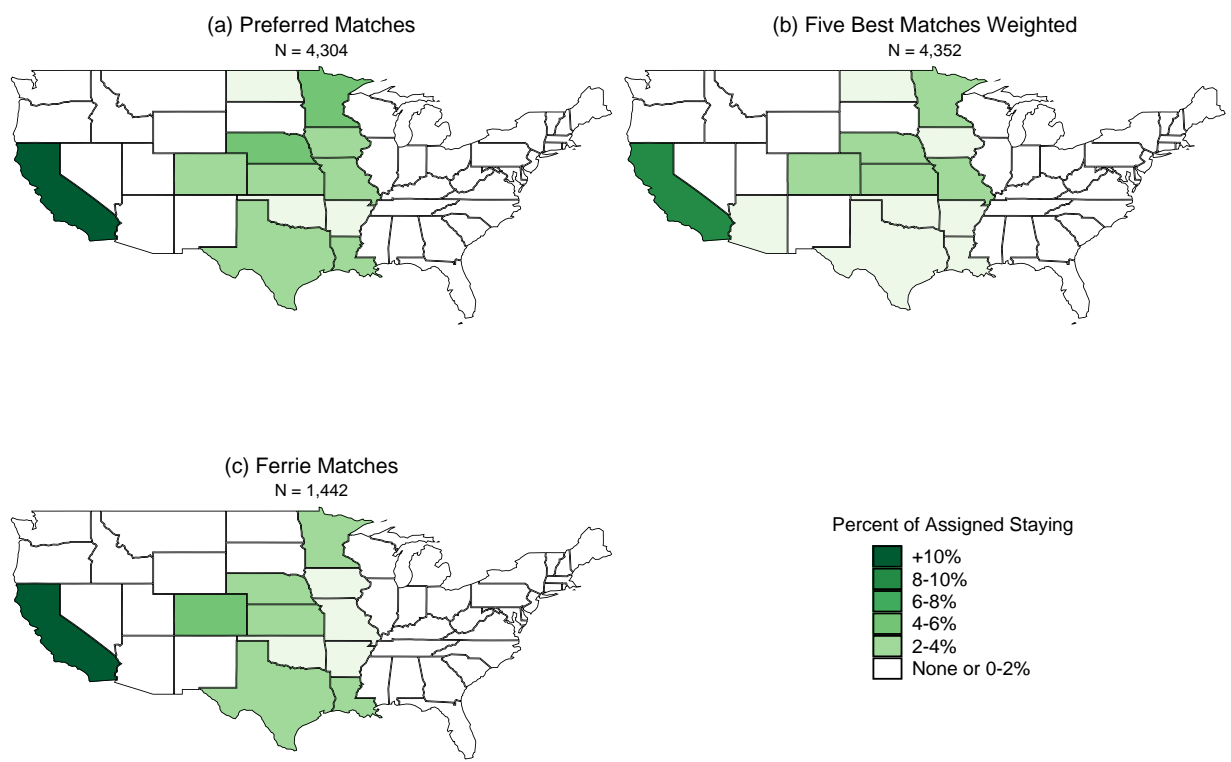
Notes: Panel (a) shows the share of male Galveston participants, who would have been 15 to 44 in 1910, assigned to each state upon their arrival in Galveston between 1908 and 1914. Panel (b) shows the share assigned to each state upon arrival but uses the sample that can be matched to the 1920 census using our preferred method.

Figure 3: State Location as of 1920 Census



Notes: Panels (a) to (c) show the share of prime-age Galveston immigrants located in each state as of the 1920 census, by matching method. See the text for more information on our matching procedures. Panel (d) shows the share of non-Galveston Russian-born men, aged 25 to 54 in 1920 and arriving in the U.S. between 1908 and 1914, that are located in each state as of the 1920 census.

Figure 4: Percent of Galveston Men that Remain in Their Assigned State as of 1920



Notes: This figure shows the share of Galveston immigrants who remain in their assigned state as of the 1920 census. We do not compute shares for states with 5 or fewer assignments or east of the Mississippi.

Table 1: Summary Statistics: Full Data

	Port of Galveston		Comparable 1920 Immigrants	
	(1) Russian Hebrews	(2) Other Arrivals	(3) Russians	(4) Non-Russians
Age	25.6 (0.1)	27.6 (0.1)	33.5 (0.0)	34.0 (0.0)
Arrive Prior to 1910 (%)	18.3 (0.5)	43.5 (0.5)	21.7 (0.1)	23.4 (0.0)
West of Mississippi (%)	91.8 (0.4)	99.0 (0.1)	10.1 (0.1)	18.1 (0.0)
<i>Texas (%)</i>	37.4 (0.6)	70.3 (0.4)	0.4 (0.0)	0.5 (0.0)
<i>California (%)</i>	2.3 (0.2)	8.8 (0.3)	1.7 (0.0)	4.7 (0.0)
New York (%)	0.0 (0.0)	0.0 (0.0)	28.6 (0.1)	18.5 (0.0)
Pennsylvania-New Jersey (%)	0.0 (0.0)	0.1 (0.0)	18.5 (0.1)	20.6 (0.0)
Midwest (East of Mississippi) (%)	4.7 (0.3)	0.4 (0.1)	25.0 (0.1)	28.5 (0.0)
New England (%)	0.0 (0.0)	0.0 (0.0)	14.5 (0.1)	11.2 (0.0)
South (East of Mississippi) (%)	3.4 (0.2)	0.4 (0.1)	3.3 (0.0)	3.0 (0.0)
Occupational Income (1940 dollars)	1355.3 (6.0)	957.9 (5.9)	1129.1 (1.5)	1076.7 (0.6)
Professional (%)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.1 (0.0)
Farmer, Farm Laborer (%)	7.5 (0.3)	48.4 (0.5)	4.2 (0.0)	6.2 (0.0)
Managers, Proprietors (%)	7.7 (0.3)	3.5 (0.2)	7.8 (0.1)	3.3 (0.0)
Clerical Workers (%)	8.5 (0.4)	3.7 (0.2)	0.8 (0.0)	1.1 (0.0)
Sales Workers (%)	4.3 (0.3)	0.6 (0.1)	3.4 (0.0)	1.6 (0.0)
Craftsmen (%)	46.3 (0.6)	19.9 (0.4)	19.3 (0.1)	16.8 (0.0)
Operative Workers (%)	14.8 (0.5)	9.7 (0.3)	21.1 (0.1)	20.1 (0.0)
Service Workers (%)	2.2 (0.2)	2.6 (0.2)	2.6 (0.0)	4.4 (0.0)
General Laborer (%)	5.0 (0.3)	5.6 (0.2)	19.4 (0.1)	26.2 (0.0)
No Occupation (%)	3.7 (0.2)	6.1 (0.2)	21.5 (0.1)	20.1 (0.0)
N	5,911	10,560	234,618	1,147,043

Notes: Columns (1) shows our unmatched sample of prime-age male Galveston participants. Columns (2) shows all prime-age male non-Russian arrivals to the port of Galveston over the same period. Columns (3) and (4) compare these samples to the total Russian-born and non-Russian immigrant populations arriving from 1908 to 1914 and meeting the same sample restrictions.

Table 2: Comparison of Match Rates and Synthetic Validation

	Year	Actual Match Rate (%)	Synthetic Match Rate (%)	Synthetic Type I Error (%)
Preferred	1910	71.3	88.4	34.5
	1920	72.8	87.7	35.3
	1930	69.5	87.0	34.6
Ferrie	1910	25.1	48.0	23.4
	1920	24.4	46.8	22.2
	1930	23.3	47.8	23.5
Five Best Weighted	1910	71.9	89.2	63.3
	1920	73.4	88.8	64.0
	1930	70.2	88.1	63.0

Notes: See the text for details on our matching procedures. See Appendix B for information on the synthetic tests.

Table 3: Summary Statistics: Matched Data

	(1) Matched Sample at Arrival	(2) 1910 Matches	(3) 1920 Matches	(4) 1930 Matches	(5) 1940 Matches
Age	25.3 (0.1)	27.2 (0.2)	34.7 (0.1)	44.5 (0.1)	55.3 (0.1)
Arrive Prior to 1910 (%)	18.2 (0.6)	54.1 (1.3)	18.2 (0.6)	17.9 (0.6)	18.2 (0.6)
Mother Tongue: Hebrew (%)	. (.)	51.8 (1.3)	64.3 (0.7)	68.8 (0.7)	. (.)
Mother Tongue: Russian (%)	. (.)	8.1 (0.7)	23.9 (0.6)	20.6 (0.6)	. (.)
Stay Assigned State (%)	. (.)	3.2 (0.5)	3.5 (0.3)	3.0 (0.3)	2.0 (0.2)
Stay Assigned Division (%)	. (.)	4.6 (0.5)	6.3 (0.4)	5.3 (0.3)	4.5 (0.3)
West of Mississippi (%)	91.6 (0.4)	9.5 (0.8)	11.6 (0.5)	12.4 (0.5)	18.5 (0.6)
<i>Texas (%)</i>	36.8 (0.7)	1.4 (0.3)	1.3 (0.2)	1.1 (0.2)	0.4 (0.1)
<i>California (%)</i>	2.4 (0.2)	0.9 (0.3)	1.6 (0.2)	3.6 (0.3)	9.3 (0.4)
New York (%)	0.0 (0.0)	45.5 (1.3)	37.9 (0.7)	37.7 (0.8)	33.7 (0.7)
Pennsylvania-New Jersey (%)	0.0 (0.0)	16.5 (1.0)	14.5 (0.5)	14.8 (0.6)	12.6 (0.5)
Midwest (East of Mississippi) (%)	4.6 (0.3)	17.3 (1.0)	22.6 (0.6)	22.0 (0.6)	20.2 (0.6)
New England (%)	0.0 (0.0)	8.7 (0.7)	9.6 (0.4)	10.1 (0.5)	11.0 (0.5)
South (East of Mississippi) (%)	3.7 (0.3)	2.5 (0.4)	3.8 (0.3)	3.0 (0.3)	4.1 (0.3)
Married (%)	. (.)	47.7 (1.3)	78.0 (0.6)	88.6 (0.5)	83.9 (0.5)
Naturalized (%)	. (.)	0.6 (0.2)	49.1 (0.8)	78.4 (0.6)	86.1 (0.5)
Literate (%)	. (.)	76.9 (1.1)	83.8 (0.6)	89.0 (0.5)	. (.)
Owns Home (%)	. (.)	8.4 (0.7)	17.1 (0.6)	35.1 (0.7)	35.9 (0.7)
Occupational Income (1940 dollars)	1357.8 (7.2)	1245.5 (12.0)	1171.1 (11.9)	1266.4 (13.1)	1396.1 (12.2)
Self-Employed (%)	. (.)	9.0 (0.7)	20.8 (0.6)	28.2 (0.7)	34.0 (0.7)
Works for Wages (%)	. (.)	86.1 (0.9)	55.9 (0.8)	49.5 (0.8)	46.1 (0.7)
Farm Status (%)	. (.)	1.9 (0.4)	1.8 (0.2)	2.9 (0.3)	3.7 (0.3)
Professional (%)	0.0 (0.0)	0.0 (0.0)	0.1 (0.0)	0.1 (0.0)	0.2 (0.1)
Farmer, Farm Laborer (%)	7.3 (0.4)	1.7 (0.3)	2.1 (0.2)	3.0 (0.3)	3.5 (0.3)
Managers, Proprietors (%)	8.1 (0.4)	2.4 (0.4)	11.8 (0.5)	19.2 (0.6)	21.1 (0.6)
Clerical Workers (%)	8.6 (0.4)	1.4 (0.3)	0.8 (0.1)	0.8 (0.1)	2.3 (0.2)
Sales Workers (%)	4.3 (0.3)	4.5 (0.5)	5.3 (0.3)	6.1 (0.4)	9.8 (0.4)
Craftsmen (%)	46.1 (0.8)	23.9 (1.1)	21.6 (0.6)	21.2 (0.6)	18.5 (0.6)
Operative Workers (%)	14.9 (0.5)	29.5 (1.2)	20.8 (0.6)	16.1 (0.6)	15.3 (0.5)
Service Workers (%)	2.1 (0.2)	1.9 (0.4)	2.1 (0.2)	3.3 (0.3)	5.1 (0.3)
General Laborer (%)	4.8 (0.3)	27.4 (1.2)	11.7 (0.5)	7.4 (0.4)	6.6 (0.4)
No Occupation (%)	3.9 (0.3)	7.5 (0.7)	23.7 (0.6)	22.9 (0.7)	17.5 (0.6)
N	4,304	1,487	4,304	4,109	4,699

Notes: Column (1) shows the arrival characteristics of prime-age male Galveston participants for those who were matched to the 1920 census. Columns (2) to (5) show census variables for this sample conditional on being matched to each census.

Table 4: Determinants of 1920 State Location

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Assigned	1.101*** (0.104) [0.0188]	0.724*** (0.089) [0.0128]	1.233*** (0.093) [0.0213]	1.209*** (0.092) [0.0206]	0.945*** (0.097) [0.0162]	1.137*** (0.097) [0.0194]	1.178*** (0.094) [0.0204]	1.234*** (0.093) [0.0213]	1.161*** (0.097) [0.0198]	1.132*** (0.102) [0.0193]	1.086*** (0.103) [0.0185]
Log(Pop.) 1910		0.419*** (0.049) [0.0074]	0.002 (0.051) [0.0000]	0.016 (0.050) [0.0003]	-0.041 (0.053) [-0.0007]	-1.480*** (0.326) [-0.0253]	-0.213*** (0.078) [-0.0037]	0.002 (0.051) [0.0000]	0.012 (0.091) [0.0002]	-0.453 (0.429) [-0.0077]	-0.566 (0.435) [-0.0097]
Log(Immig.) 1910		1.117*** (0.039) [0.0198]	0.287*** (0.053) [0.0050]	0.292*** (0.050) [0.0050]	0.011 (0.052) [0.0002]	0.125 (0.177) [0.0021]	0.377*** (0.059) [0.0065]	0.286*** (0.053) [0.0049]	0.178*** (0.065) [0.0030]	0.045 (0.210) [0.0008]	-0.006 (0.217) [-0.0001]
Log(Russian Immig.) 1910			0.615*** (0.037) [0.0107]	0.558*** (0.036) [0.0095]	0.217*** (0.043) [0.0037]	0.529*** (0.045) [0.0090]	0.626*** (0.036) [0.0108]	0.616*** (0.037) [0.0107]	0.715*** (0.048) [0.0122]	0.649*** (0.058) [0.0111]	0.362*** (0.074) [0.0062]
Log(Jewish+1) 1906			0.177*** (0.025) [0.0031]	0.141*** (0.023) [0.0024]	0.006 (0.031) [0.0001]	0.178*** (0.030) [0.0030]	0.256*** (0.034) [0.0044]	0.177*** (0.025) [0.0031]	0.141*** (0.043) [0.0024]	0.107** (0.050) [0.0018]	0.068 (0.056) [0.0012]
Located 1920, Ship (%)				0.010*** (0.002) [0.0002]							0.001 (0.003) [0.0000]
D(No Non-Fam. Located) 1920					0.327 (0.507) [0.0056]						0.504 (0.543) [0.0086]
Log(Num. Non-Fam. Located 1920)					0.810*** (0.055) [0.0139]						0.653*** (0.084) [0.0111]
Log(Male Immig.), Cohort (10 Yr.)						0.008 (0.215) [0.0001]				-0.545 (0.336) [-0.0093]	-0.185 (0.326) [-0.0031]
Log(Male US Born), Cohort (10 Yr.)						-0.271 (0.562) [-0.0046]				-0.654 (0.747) [-0.0111]	0.653 (0.785) [0.0111]
Literate (%), Male Immig, Cohort (10 Yr.)					0.092*** (0.015) [0.0016]					0.057*** (0.019) [0.0010]	0.020 (0.021) [0.0003]
Literate (%), Male US Born, Cohort (10 Yr.)					-0.049 (0.039) [-0.0008]					-0.071* (0.042) [-0.0012]	-0.086** (0.043) [-0.0015]
Log(Female Immig.), Cohort (10 Yr.)					0.332** (0.151) [0.0057]					0.722*** (0.198) [0.0123]	0.267 (0.194) [0.0046]
Log(Female US Born), Cohort (10 Yr.)					1.634*** (0.597) [0.0279]					1.193 (0.741) [0.0203]	-0.192 (0.802) [-0.0033]
Literate (%), Female Immig, Cohort (10 Yr.)					-0.053*** (0.013) [-0.0009]					-0.045*** (0.016) [-0.0008]	-0.003 (0.018) [-0.0001]
Literate (%), Female US Born, Cohort (10 Yr.)					0.050 (0.037) [0.0009]					0.058 (0.039) [0.0010]	0.042 (0.038) [0.0007]
Log(State Mean Occscore), 1910							-1.127*** (0.303) [-0.0195]			2.440 (1.583) [0.0416]	-0.115 (1.579) [-0.0020]
Workers with own 2-Dig. Occ. (%), 1910								-0.011 (0.018) [-0.0002]		0.018 (0.018) [0.0001]	0.018 (0.018) [-0.0001]
Farmers (%), 1910									-0.126*** (0.036) [0.0022]	-0.056 (0.043) [0.0021]	-0.056 (0.047) [-0.0010]
Farm Laborers (%), 1910									0.078*** (0.026) [0.0013]	-0.021 (0.043) [-0.0004]	0.001 (0.043) [0.0000]
Professionals (%), 1910									1.087*** (0.194) [0.0185]	1.144*** (0.228) [0.0195]	0.568** (0.242) [0.0097]
Managers (%), 1910									-1.251*** (0.153) [-0.0213]	-0.921*** (0.242) [-0.0157]	-0.259 (0.239) [-0.0044]
Clerical, Sales, and Service (%), 1910									0.031 (0.035) [0.0005]	-0.142** (0.057) [-0.0024]	-0.143*** (0.054) [-0.0024]
Operatives (%), 1910									-0.192*** (0.022) [-0.0033]	-0.212*** (0.040) [-0.0036]	-0.060 (0.043) [-0.0010]
Craftsmen (%), 1910									0.276*** (0.049) [0.0047]	0.116 (0.081) [0.0020]	0.138* (0.081) [0.0024]
χ^2 (LR test): Population			8825.0***	3133.2***	31.3***	174.4***	5367.5***	8823.4***	2411.9***	133.2***	26.0***
χ^2 (LR test): Cohort						140.2***				36.5***	13.4*
χ^2 (LR test): Occupation										58.0***	20.8**
χ^2 (LR test): Galveston Immigrants											61.1***
Pseudo-R ²	0.382	0.359	0.372	0.373	0.379	0.376	0.372	0.372	0.377	0.378	0.380
N	4,304	4,304	4,304	4,304	4,304	4,304	4,304	4,304	4,304	4,304	4,304

Notes: This table implements McFadden's (1974) alternative-specific conditional logit model. Column (1) includes state intercepts to absorb state-specific variation. *Located 1920, Ship (%)* is the share of non-family shipmates who lived in a state in 1920. *Log(Num. Non-Fam. Located 1920)* is the log-number of non-family members residing in a state in 1920. In order to take logs, this variable is imputed as zero when a participant has no non-family members in a state, and we include the dummy, *D(No Non-Fam. Located 1920)*, for these cases. Cohort variables are measured using data on members of a participant's 10-year birth cohort who reside in a state. Marginal effects are calculated as the mean of each state-specific marginal effect (see footnote 27). Standard errors are in parentheses and average marginal effects are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

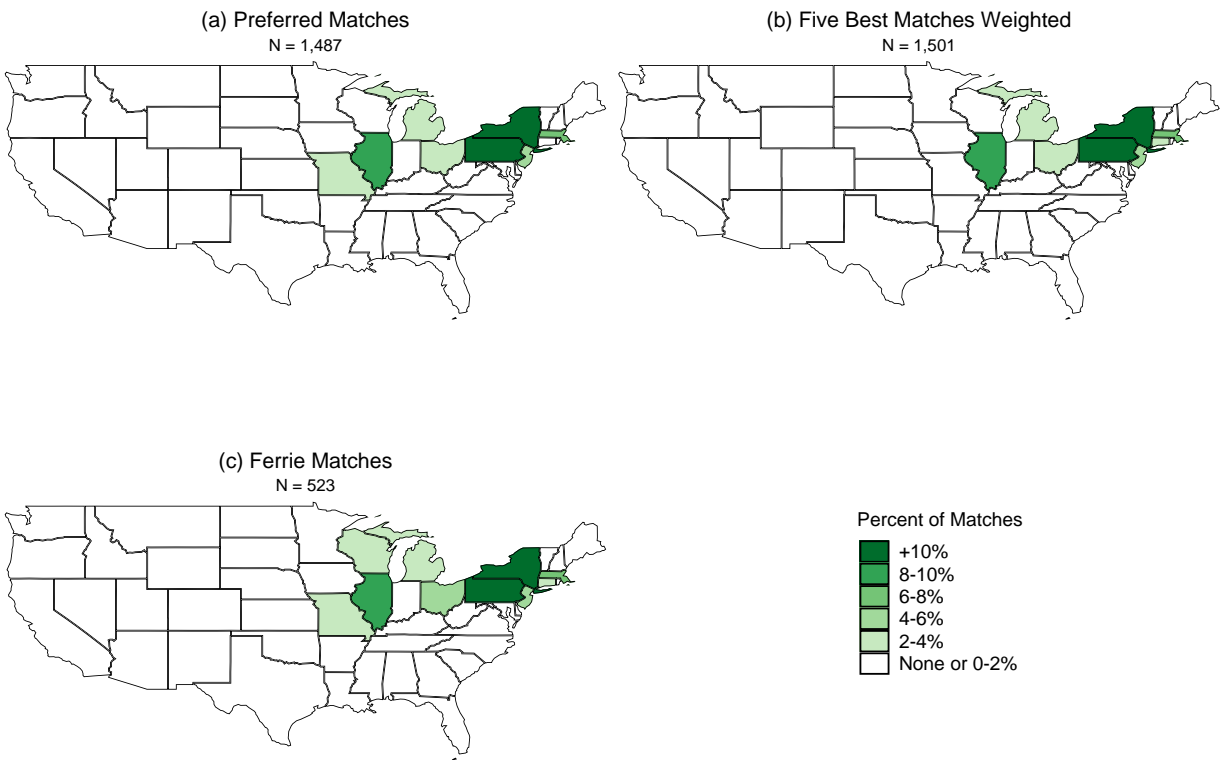
Table 5: Determinants of 1920 State Location, by Occupational Income Tercile

	Tercile 1				Tercile 2				Tercile 3			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Assigned	1.047*** (0.157)	1.068*** (0.157)	1.035*** (0.164)	0.977*** (0.166)	1.254*** (0.154)	1.266*** (0.155)	1.332*** (0.161)	1.292*** (0.164)	1.025*** (0.213)	1.029*** (0.213)	1.009*** (0.222)	0.940*** (0.223)
Log(Pop.) 1910	-1.552*** (0.512)	-1.479*** (0.512)	-0.660 (0.618)	-0.379 (0.675)	-1.509*** (0.534)	-1.425*** (0.549)	-0.885 (0.681)	-0.890 (0.746)	-1.463*** (0.692)	-1.468** (0.696)	-0.547 (0.847)	-0.303 (0.913)
Log(Immig.) 1910	0.047 (0.281)	0.183 (0.269)	-0.050 (0.295)	-0.032 (0.337)	-0.355 (0.312)	-0.272 (0.305)	-0.259 (0.323)	-0.261 (0.350)	0.535 (0.391)	0.569 (0.373)	0.341 (0.425)	0.416 (0.482)
Log(Russian Immig.) 1910	0.499*** (0.069)	0.498*** (0.070)	0.570*** (0.088)	0.312*** (0.115)	0.481*** (0.072)	0.480*** (0.073)	0.628*** (0.098)	0.312** (0.127)	0.665*** (0.100)	0.667*** (0.100)	0.858*** (0.117)	0.527*** (0.150)
Log(Jewish+1) 1906	0.212*** (0.060)	0.144*** (0.047)	0.106 (0.077)	0.084 (0.090)	0.210*** (0.064)	0.164*** (0.052)	0.152* (0.084)	0.075 (0.090)	0.261*** (0.082)	0.250*** (0.062)	0.093 (0.109)	0.044 (0.118)
Located 1920, Ship (%)				-0.005 (0.005)				0.008 (0.005)				-0.000 (0.007)
D(No Non-Fam. Located) 1920					1.174* (0.678)			-0.213 (1.074)				-12.272*** (4.28)
Log(Num. Non-Fam. Located 1920)				0.666*** (0.145)				0.572*** (0.129)				0.745*** (0.159)
Log(Male Immig.), Cohort (10 Yr.)	0.346 (0.355)	0.136 (0.334)	0.070 (0.405)	0.249 (0.497)	0.276 (0.377)	0.166 (0.363)	-0.515 (0.459)	-0.365 (0.529)	-0.289 (0.485)	-0.328 (0.451)	-0.660 (0.601)	-0.620 (0.737)
Log(Male US Born), Cohort (10 Yr.)	-2.325** (0.946)	-2.097** (0.911)	-2.615** (1.202)	-1.949 (1.502)	0.335 (0.928)	0.364 (1.054)	1.109 (1.144)	1.910* (1.144)	1.152 (1.132)	1.200 (1.112)	1.424 (1.343)	2.213 (1.533)
Literate (%), Male Immig. Cohort (10 Yr.)	0.105*** (0.022)	0.106*** (0.022)	0.072** (0.030)	0.034 (0.033)	0.096*** (0.025)	0.096*** (0.025)	0.103*** (0.033)	0.075** (0.036)	0.060* (0.032)	0.060* (0.032)	-0.027 (0.039)	-0.074* (0.042)
Literate (%), Male US Born, Cohort (10 Yr.)	-0.016 (0.060)	-0.011 (0.060)	-0.013 (0.063)	-0.031 (0.064)	-0.067 (0.062)	-0.064 (0.062)	-0.148** (0.069)	-0.144** (0.071)	-0.085 (0.088)	-0.083 (0.087)	-0.102 (0.086)	-0.117 (0.090)
Log(Female Immig.), Cohort (10 Yr.)	0.173 (0.239)	0.171 (0.233)	0.331 (0.256)	-0.053 (0.303)	0.569** (0.257)	0.546** (0.259)	0.988*** (0.289)	0.694** (0.321)	0.204 (0.324)	0.198 (0.323)	0.483 (0.372)	0.166 (0.417)
Log(Female US Born), Cohort (10 Yr.)	3.619*** (1.000)	3.487*** (0.992)	3.261** (1.266)	2.139 (1.494)	1.063 (0.942)	1.062 (0.951)	-0.064 (1.065)	-0.951 (1.151)	-0.047 (1.196)	-0.065 (1.192)	-0.955 (1.461)	-2.145 (1.627)
Literate (%), Female Immig. Cohort (10 Yr.)	-0.061*** (0.020)	-0.065*** (0.020)	-0.052** (0.024)	-0.015 (0.026)	-0.061*** (0.022)	-0.063*** (0.022)	-0.077*** (0.028)	-0.041 (0.030)	-0.012 (0.031)	-0.012 (0.031)	0.019 (0.036)	0.067* (0.037)
Literate (%), Female US Born, Cohort (10 Yr.)	0.027 (0.056)	0.023 (0.056)	0.025 (0.057)	0.015 (0.054)	0.086 (0.059)	0.083 (0.059)	0.124* (0.065)	0.104 (0.064)	0.037 (0.083)	0.036 (0.082)	0.035 (0.076)	0.017 (0.074)
Log(State Mean Occscore), 1910	-1.053* (0.567)			0.060 (2.564)	-0.631 (0.549)			-0.430 (2.456)	-0.181 (0.730)			-0.027 (3.524)
Workers with own 2-Dig. Occ. (%), 1910		-0.015 (0.019)		-0.009 (0.019)		0.063 (0.136)		0.084 (0.133)		-0.058 (0.268)		-0.080 (0.275)
Farmers (%), 1910			-0.134** (0.066)	-0.019 (0.072)			-0.118* (0.069)	-0.083 (0.074)			-0.182* (0.096)	-0.084 (0.109)
Farm Laborers (%), 1910				-0.0023 (0.010)	-0.0003 (0.055)			-0.0020 (0.060)	-0.0014 (0.071)		-0.0031 (0.074)	-0.0014 (0.085)
Professionals (%), 1910			0.960*** (0.332)	0.342 (0.374)			1.094*** (0.361)	0.416 (0.389)			1.986*** (0.491)	1.295** (0.556)
Managers (%), 1910				-0.840** (0.366)	-0.359 (0.380)			-0.232 (0.379)	0.345 (0.407)		-1.578*** (0.494)	-0.915* (0.499)
Clerical, Sales, and Service (%), 1910					-0.0143 (-0.0061)				0.0058 (0.0058)			-0.0156 (-0.0156)
Operatives (%), 1910				-0.051 (0.086)	-0.033 (0.085)			-0.283*** (0.095)	-0.311*** (0.093)		-0.090 (0.118)	-0.116 (0.110)
Craftsmen (%), 1910									0.004 (0.072)		-0.268*** (0.065)	-0.151* (0.090)
									0.0001 (0.0001)			-0.0026 (-0.0026)
									0.039 (0.129)		0.425*** (0.130)	0.350** (0.164)
									0.0007 (0.0007)			0.0060 (0.0060)
χ^2 (LR test): Population	67.9***	61.3***	46.0***	8.0*	59.0***	55.9***	50.3***	8.7*	56.1***	57.6***	55.4***	12.5**
χ^2 (LR test): Cohort	59.7***	65.2***	17.2**	4.7	51.3***	46.8***	34.3***	19.1**	41.5***	43.0***	6.2	7.3
χ^2 (LR test): Occupation				5.7				17.8**				12.0
χ^2 (LR test): Galveston Immigrants				22.4***				24.0***				1959.7***
Pseudo-R ²	0.379	0.379	0.380	0.383	0.383	0.383	0.385	0.387	0.363	0.363	0.367	0.371
N	1,806	1,806	1,806	1,806	1,564	1,564	1,564	1,564	934	934	934	934

Notes: This table splits the sample from Table 4 into terciles of occupational income at arrival. See notes to Table 4 for variable definitions. Standard errors are in parentheses and average marginal effects are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

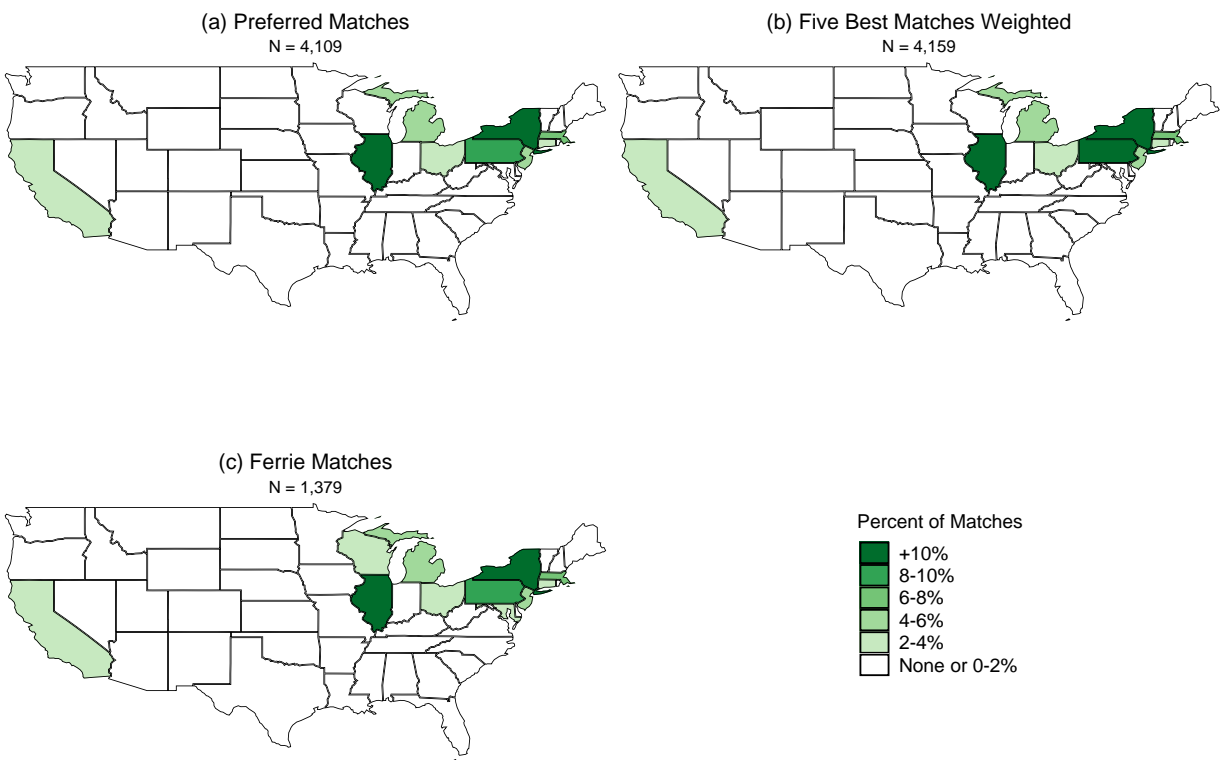
Appendix

Figure A1: State Location as of 1910 Census



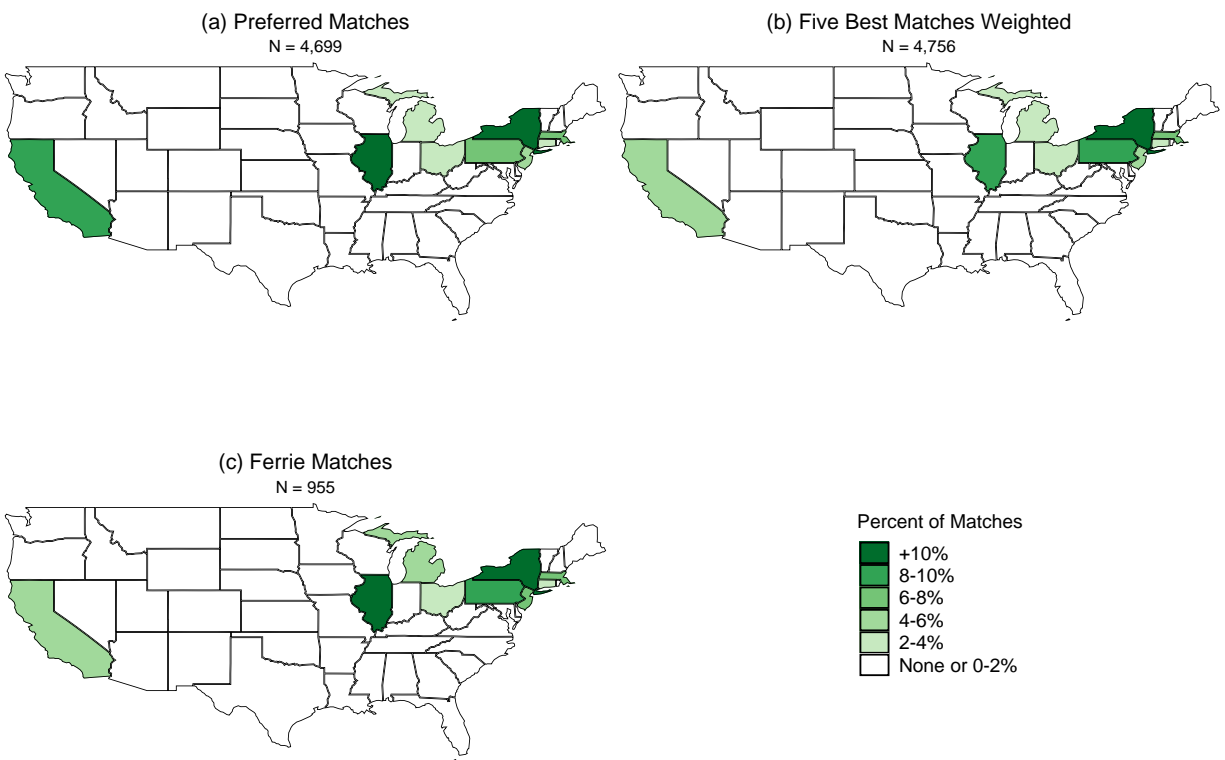
Notes: See notes to Figure 3.

Figure A2: State Location as of 1930 Census



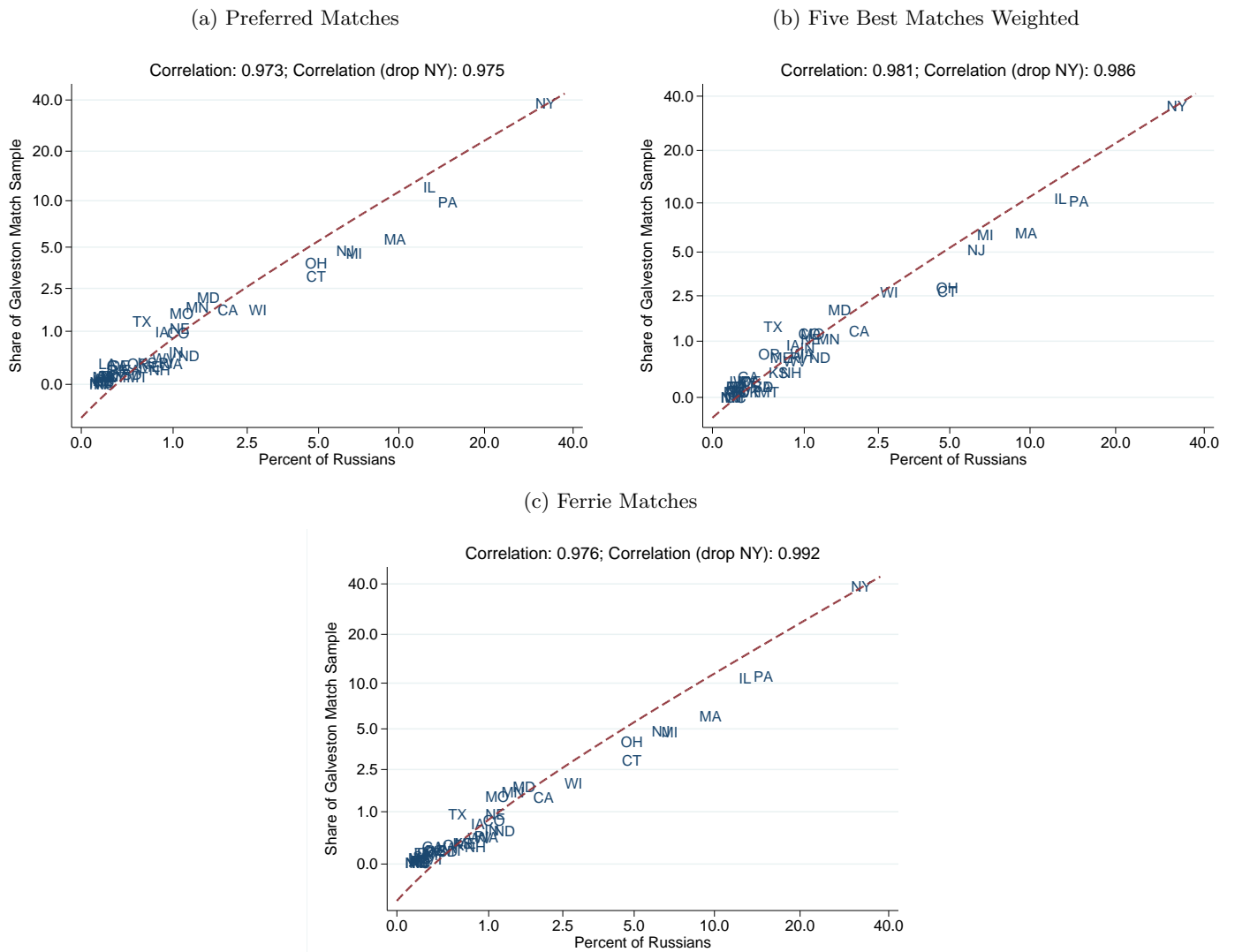
Notes: See notes to Figure 3.

Figure A3: State Location as of 1940 Census



Notes: See notes to Figure 3.

Figure A4: Comparison of Locations of Galveston Matches with Comparable Russian Immigrants in 1920



Notes: This figure plots the share of Galveston men in the 1920 census located in a state against the share of non-Galveston Russian-born men of the same age and arrival cohort in a state, as of the 1920 census. Axes are on log-scales. Correlations between state-level shares are displayed above the figure, including and excluding New York state.

Table A1: Match Parameters

	Values	Lower Limit	Upper Limit
$\beta_{Constant}$	1111.49	0	3000
$\beta_{BornRussia}$	644.13	0	1000
β_{Age}	-110.00	-1000	0
$\beta_{YearArrive}$	-390.18	-1000	0
β_{Last}	807.12	0	1000
β_{First}	295.07	0	1000
$\beta_{Last \times First}$	847.11	0	1000
$\beta_{SpeakRussian}$	226.73	0	1000
$\beta_{SpeakHebrew}$	329.22	0	1000
γ_{Arr}	3.26	1	5
γ_{Age}	1.10	1	5
δ_{Age}	12	0	15

Notes: See Appendix A.

Table A2: Comparison of 1920 Location Results across Matching Methods

	(1)	(2)	(3)
	Preferred	Ferrie	Five Best Matches Weighted
Assigned	0.946*** (0.097) [0.0162]	1.003*** (0.171) [0.0176]	0.655*** (0.068) [0.0111]
Log(Pop.) 1910	-0.039 (0.053) [-0.0007]	-0.093 (0.085) [-0.0016]	-0.026 (0.031) [-0.0004]
Log(Immig.) 1910	0.017 (0.052) [0.0003]	0.144 (0.090) [0.0025]	-0.007 (0.034) [-0.0001]
Log(Russian Immig.) 1910	0.217*** (0.042) [0.0037]	0.193*** (0.075) [0.0034]	0.140*** (0.030) [0.0024]
Log(Jewish+1) 1906	0.004 (0.030) [0.0001]	0.036 (0.042) [0.0006]	-0.005 (0.018) [-0.0001]
D(No Non-Fam. Located) 1920	0.264 (0.515) [0.0045]	0.986*** (0.280) [0.0173]	
Located 1920, Ship (%)	0.001 (0.002) [0.0000]	-0.004 (0.003) [-0.0001]	-0.000 (0.001) [-0.0000]
Log(Num. Non-Fam. Located 1920)	0.795*** (0.061) [0.0136]	0.753*** (0.100) [0.0132]	0.895*** (0.046) [0.0152]
χ^2 (LR test): Population	31.8***	13.0**	24.8***
χ^2 (LR test): Galveston Immigrants	219.8***	75.3***	503.8***
Pseudo-R ²	0.379	0.350	0.383
N	4,304	1,442	4,352

Notes: See notes to Table 4. Standard errors are in parentheses and average marginal effects are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.