A Surname-Based Measure of Ethnicity, and Its Applications to Studies of Ethnic Segregation: The Early Twentieth Century U.S.[*]

Dafeng Xu[†]

Abstract

Many studies consider ethnicity as an equivalent term for the country of birth. This could neglect ethnic heterogeneity within the sending country. Focusing on German-born, Polish-born, and Russian-born immigrants in the 1920 and 1930 U.S. census, I propose an ethnicity variable constructed based on the linguistic origin of the surname using an artificial intelligence algorithm. Employing this ethnicity variable, I study ethnic segregation within each immigrant group defined based on the country of birth. Results suggest the degree of within-group ethnic segregation was high. Specifically, ethnic majorities within each immigrant group generally resided in areas with significantly more compatriots.

Keywords: immigration, surname, ethnicity, segregation, early twentieth century U.S.

JEL Classification: J1, N3, R1

[*] I thank Fran Blau, Nancy Brooks, Jack DeWaard, Marina Gorsuch, Dave Hacker, Larry Kahn, Dan Lichter, Mao-Mei Liu, Matt Nelson, Evan Roberts, Steve Ruggles, Matt Sobek, Tony Smith, Yannay Spitzer, Eddie Telles, Dave Van Riper, Zach Ward, Rob Warren, Klaus Zimmermann, Ariell Zimran, and seminar participants at Cornell University, University of Minnesota, and the 2018 Population Association America (PAA) meeting for their valuable suggestions. I also receive helpful comments from Ran Abramitzky on a highly related project. I thank Joe Grover, Kai Hong, and Jonathan Schroeder for their discussions on data and methodologies in this paper.

[†] 225 19th Ave. S., University of Minnesota, Minneapolis, MN 55455. Email: dafengxu@umn.edu.

1 Introduction

Many studies in labor, population and regional economics consider *ethnicity*, or *ethnic origin*, as the equivalent term for *country of birth* (e.g., Fairlie and Meyer, 1996; Bleakley and Chin, 2004). Compared to *ethnicity*, *country of birth* is a more political and geographical measure of *origin*, while *ethnicity* is more relevant to genetics and culture. Hence, for an immigrant, his country of birth and ethnicity might be different. This was especially true in the U.S. in the early twentieth century, when most U.S. immigrants were from Europe, and many European countries had higher degrees of ethnic diversity. For example, an immigrant born in Germany might actually be of Polish ethnicity, and belonged to the Polish ethnic minority group back in Germany. This suggests that there should be within-group ethnic differences if immigrants are classified by the country of birth.

Cutler et al. (1998) point out that immigrant segregation was high in the early twentieth century U.S. Unsurprisingly, some immigrant groups (e.g., German) were less segregated and spatially more assimilated than other groups (e.g., Polish, Russian). The possible within-group ethnic differences lead to the following questions: if there were indeed such differences, whether (and how) ethnic majorities and minorities were spatially segregated. Hence, classifying ethnicity is not only methodologically interesting, but also useful for understanding ethnic segregation, which is an important topic in urban and labor economics.

The specific context of this paper lies in the early twentieth century U.S. In the 1910s and 1920s, first-generation immigrants made up nearly 15% of the U.S. population. Most immigrants were born in Europe, and Germany, Poland, and Russia were among the top sending countries of immigrants. Compared with other sending countries, these countries had higher degrees of ethnic and cultural diversity, and there were possibly ethnic differences within each immigrant group defined based on the country of birth.

The outline of the empirical framework is summarized as follows: I first construct an auxiliary variable that indicates the linguistic origin of the surname for each German-born, Polish-born, and Russian-born immigrant in the 1920 and 1930 U.S. census. Using this variable to proxy for ethnicity, I focus on each immigrant group defined based on the country of birth, and examine within-group ethnic segregation (e.g., segregation between German-born immigrants of German and Russian ethnicity). I also study the structure of segregation by examining immigrant enclave residence (e.g., whether German-born immigrants of Russian ethnicity were less likely to live in German enclaves).

To start, I first construct the surname-based ethnicity variable. The linguistic origin of the surname is highly related to ethnicity, as pointed out by research in human biology and ethnography that the surname origin reflects both genetic and cultural transmission (Waters, 1989; Chibnik, 1991; Guglielmino et al., 2000; Schramm et al., 2012). This idea has been used in many disciplines, such as epidemiology (Razum et al., 2001), geography (Mateos, 2007), demography (Monasterio, 2017), as well as economics (Foley and Kerr, 2013). The Census Bureau also identifies the Hispanic origin based on the surname (Ruggles, 2017).

A straightforward strategy of classifying surname-based ethnicity is to match surnames in census data with a dictionary of German, Polish, and Russian surnames, and a dictionary of Anglicized names (since some immigrants Americanized their names, e.g., Abramitzky et al., 2016; Biavaschi et al., 2017). This, however, would leave many surnames in census data unmatched, because (a) many immigrants moderately—but not fully—converted their names (e.g., from *Eisenhauer* to *Eisenhower*); (b) there are misspellings and transcription errors in digitized census data (e.g., *Ivanov* is digitized as *Lvanov*). To further study this, I use an artificial intelligence algorithm—more specifically, the naïve Bayes classifier (Rish, 2005)—to classify the linguistic origin of the surname using a large training dataset of ethnicity-specific surnames. The basic idea is to calculate the probability of "occurrence" for each string (in surnames) in different languages based on training data, and then use the *a priori* probabilities to predict the linguistic origin of surnames of unknown ethnicity. Before implementing this algorithm in real census data, I test its performance in a validation dataset, in which reliable answers of individuals' ethnicity are known. Classification results suggest that this algorithm performs well in the validation dataset.

I then employ this algorithm in the 1920 and 1930 census, focusing on first-generation immigrants who reported Germany, Poland, and Russia as the country of birth. At that time, there were only four non-Anglophone countries—Italy, Germany, Poland, and Russia—that sent more than one million immigrants, and Germany, Poland, and Russia had high degrees of ethnic diversity. I only study the 1920 and 1930 census because (a) many countries (e.g., Poland) were not independent before World War I, and thus census takers might not record them as immigrants' country of birth prior to 1920;[1] (b) in 1940, several relevant questions (e.g., year of immigration) were not surveyed. Classification results indeed show huge within-group differences in the linguistic origin of surnames, suggesting ethnic differences in each immigrant group defined based on the country of birth.

Subsequently, I turn to the next part of the empirical analysis: using surname-based ethnicity to study ethnic segregation within the immigrant group defined based on the country of birth. I first calculate the county-level dissimilarity index of segregation among surname-based ethnic populations within each immigrant group. Results suggest high degrees of ethnic segregation in all three groups, and such segregation patterns were even comparable to immigrant-native segregation in the early twentieth century U.S.

To explore the "micro-structure" of ethnic segregation, I further study ethnic enclave residence at the enumeration district (ED) level. In each immigrant group, I regress the size of the country-of-birth enclave on surname-based ethnicity, and observe that ethnic majorities resided in significantly larger country-of-birth enclaves in all three immigrant groups. Although the results do not necessarily reflect causality, I discuss several statistical issues: I argue that measurement errors should not drive the findings, and the results are robust when I include county fixed effects, focus on the urban and rural sample separately, and control for ED-level labor market characteristics.

The empirical conclusion of this paper indicates that (surname-based) ethnicity, along with the country of birth, could play an important role in determining immigrants' settle-

[1] An example related to this paper is that there were almost no immigrants (approximately 30,000 only) who reported Poland as the country of birth in the 1910 U.S. census.

ment patterns. Considering all immigrants born in a specific country as a demographically homogeneous population might neglect huge within-group ethnic heterogeneity, which was especially true in the historical context of this paper. By applying surname-based ethnicity in studies of ethnic segregation, this paper adds to the literature of economic history, population economics, and urban economics.

In the rest of this paper, I first discuss the historical background in Section 2. I then introduce the classification algorithm in Section 3. I also examine the performance of the algorithm in a validation dataset. In Section 4, I turn to actual census data and analyze empirical strategies. In Section 5 I report results. I conclude the paper in Section 6.

2 Background

I first briefly discuss the background of this paper. The U.S. absorbed over 20 million immigrants during the age of mass migration (Abramitzky and Boustan, 2017). In the nineteenth century, most immigrants were originally from "old source countries" such as Ireland and Germany. While these countries—especially Germany—still sent many immigrants in the early twentieth century, Southern and Eastern Europe started to send an increasing number of immigrants (Haines, 2000). In 1920 and 1930, Germany (1.6 million in 1920; 1.6 million in 1930), Poland (1.1 million in 1920, 1.3 million in 1930), and Russia (1.5 million in 1920, 1.2 million in 1930) were among the top countries of origin in the U.S.

Many economic studies consider ethnicity as the equivalent term for the country of birth (e.g., Fairlie and Meyer, 1996; Bleakley and Chin, 2004). This is a reasonable assumption if immigrants from a sending country are ethnically homogeneous, which might be roughly true in the contemporary U.S. In the early twentieth century, however, many European countries had fairly high degrees of ethnic diversity, and could possible seznd both ethnic majority and (multiple) minority groups to the U.S.

In the specific context of this paper, Germany, Poland, and Russia all had high degrees of ethnic and cultural diversity prior to the 1930s, and were likely to send various ethnic groups to the U.S. Migration among three countries were not uncommon. Poland did not gain independence until 1918, and thus many Germans and Russians lived in their colonial part of Poland, and similarly, many Polish people lived in Germany and Russia. However, the high degree of ethnic diversity was not only related to Poland independence. For almost ten centuries (until the WWII), there were many German settlements in Central and Eastern Europe throughout the long historical period from the Hanseatic League and Teutonic Knights to the Kingdom of Prussia (Sammartino, 2010). In addition, prior to 1933, there was a large Jewish population in Central and Eastern Europe. In the late eighteenth century, Jews were required to culturally assimilate into the local society by, say, adopting local surnames (Stern and Rottenberg, 1998; Dubin, 1999), which did lead to Jewish emancipation nonetheless, and further encourage Jewish settlements in Central and Eastern Europe.

In each of the three sending countries studied in this paper, there was an ethnic majority group, but ethnic minority groups were still large, and individuals from both groups could move to the U.S. Moreover, immigrants from the majority group in the home country might not necessarily be ethnic majorities after arrival. For example, compared with the majority group in Russia (i.e., Russians), German ethnics and Jews had more incentives to migrate (e.g., Boustan, 2007; Fussell, 2014). This further led to ethnic differences in return migration within the immigrant group from the same country (Ward, 2017). In addition, back in Europe, some ethnic groups were more knowledgeable about immigration, which could also affected immigration to the U.S. (Haines, 2000).

3 The Classification Algorithm

In this section, I first introduce the artificial intelligence algorithm of ethnicity classification based on the linguistic origin of the surname. I then test its performance in a validation dataset of surnames, in which answers of ethnicity are known.

3.1 The Description of the Classification Algorithm

The classification problem of this paper is: given some ethnic-name dictionaries—or in the language of artificial intelligence, the training set (Rish, 2005)—how a surname of unknown ethnicity (i.e., in census data) can be correctly assigned an origin. The algorithm contains two stages. The first stage involves four sub-stages of deterministic algorithms. In the first two sub-stages I identify Americanized surnames, other European ethnicity, and Jewish ethnicity; in the second two sub-stages I identify surnames of German, Polish, and Russian ethnicity. As the deterministic algorithm in the first stage could leave many names unclassified, I employ a probabilistic algorithm in the remaining sample in the second stage.

Table 1: A List of Surnames (To Be Classified)

Number	Surname	Place of birth	Mother tongue	Number	Surname	Place of birth	Mother tongue
1	Smith	Poland	German	2	Dvorak	Germany	German
3	Bronstein	Poland	Yiddish	4	Rabinovich	Russia	Russian
5	Lisowski	Poland	Polish	6	Lvanov	Russia	Russian
7	Eisenhower	Russia	German	8	Khrushchell	Russia	Russian

In Table 1, I list eight surnames whose ethnic origins are to be classified. As an example of the algorithm implementation, I will explain how surname-based ethnic origins of these surnames are classified in each stage of the algorithm.

3.1.1 First Stage: A Deterministic Algorithm

I start with the first stage of the algorithm. The first stage of the algorithm involves four sub-stages, in which surname-based ethnicity is classified deterministically.

Sub-stage 1: I match the surnames in the sample with a dictionary of Anglicized surnames (Reaney, 2005). If an immigrant's name is matched with an Anglicized name, then ethnicity is unidentifiable because the name might have been Americanized, which was common in the early twentieth century. Similarly, I match the surnames with the dictionary of Italian, Czech, Slovak, Romanian, Hungarian and Scandinavian names (e.g., Fucilla, 1998), as these ethnics might also resided in Germany, Poland, and Russia, although these populations were relatively small.

Sub-stage 2: I identify typical cases of Jewish ethnicity. The U.S. census surveyed the mother tongue in 1920 and 1930. Based on this, any individuals speaking Yiddish or Hebrew can be identified as Jews (Sassler, 2005). Note that the 1897 Russian census indicates that 97% of all Russian Jews spoke Yiddish (Kreindler, 1985). However, some Eastern European Jews might convert to speak the local language (Corrisin, 1990). Hence, not speaking Yiddish or Hebrew does not necessarily indicate non-Jewish ethnicity. Among those who did not speak Yiddish or Hebrew, I match surnames with a dictionary of Jewish surnames (Stern and Rottenberg, 1998), and further mark surnames with typical Jewish name elements (e.g., *Rabinovich* with the prefix *Rabin-*). Note that in practice, most Jews are stil identified based on the mother tongue.

Sub-stage 3: the above two sub-stages should exclude many individuals of non-German, non-Polish, and non-Russian ethnicity. In the remaining sample, I focus on the classification among German, Polish, and Russian origin. I first match surnames with the dictionary of German (Bahlow, 2002), Polish (Hoffman, 2001), and Russian (Unbegaun, 1972) names.

Sub-stage 4: I finally classify ethnicity surnames with "typical linguistic characteristics". This identifies ethnicity of typical ethnic-specific names with moderate degrees of Americanization or transcription errors. For example, the surname *Lvanov*, a misspelled version of the Russian surname *Ivanov*, is still very likely to be of Russian origin due to its suffix *-nov*. Formally, I decompose every surname in training data into three- and four-character strings; for each string, I calculate the frequency and probability of "occurrence" by language in training data. For example, for *Ivanov* in training data, I decompose it into the following strings: *#IV*, *IVA*, *VAN*, *ANO*, *NOV*, *OV**, *#IVA*, *IVAN*, *VANO*, *ANOV*, and *NOV** (where # and * represent the beginning and end of the name). I then count the number of times the strings occur in the surname dictionary by language (e.g., the string *NOV** should appear frequently among Russian surnames). Thus, I determine a surname's ethnicity in actual census data if (a) it contains a string (denoted as γ) that appears exclusively in one ethnic-name dictionary in training data (e.g., *NOV** only appears among

Russian names), and (b) γ appears more than a threshold (say, 100 times) in the entire training dataset (e.g., *NOV** appears frequently enough). This surname's ethnicity is thus $e(\gamma)$, where $e(\gamma)$ denotes which ethnic-name dictionary γ is in. If there are two such strings γ_1 and γ_2, and $e(\gamma_1) \neq e(\gamma_2)$, I do not classify the surname in this stage.

Table 2: First-Stage Classification

Surname	Place of birth	Mother tongue	Result	Stage
Smith	Poland	German	Anglicized	First stage, sub-stage 1
Dvorak	Germany	German	Others (Czech)	First stage, sub-stage 1
Bronstein	Poland	Yiddish	Jewish (by language)	First stage, sub-stage 2
Rabinovich	Russia	Russian	Jewish (by name)	First stage, sub-stage 2
Lisowski	Poland	Polish	Polish	First stage, sub-stage 3
Lvanov	Russia	Russian	Russian	First stage, sub-stage 4
Eisenhower	Russia	German	Unclassified	Not classified in this stage
Khrushchell	Russia	Russian	Unclassified	Not classified in this stage

Table 2 present first-stage classification results, based on examples listed in Table 1. In this stage, the algorithm successfully classifies ethnicity in six cases: (a) *Smith* is an Anglicized name; (b) *Dvorak* is matched in the Czech name dictionary; (c) *Bronstein* is of Jewish ethnicity because of the mother tongue; (d) *Rabinovich* is of Jewish ethnicity because of the Jewish prefix *Rabin-*; (e) *Lisowski* is matched in the Polish name dictionary; (f) *Lvanov*, which is possibly a misspelled *Ivanov*, is still identified as of Russian ethnicity because the string *-nov* appears frequently and exclusively in the Russian name dictionary. The last two surnames, *Eisenhower* and *Khrushchell*, are not classified in this stage.

3.1.2 Second Stage: A Probabilistic Naïve Bayes Algorithm

In the first stage, I classify ethnicity using the deterministic algorithm, which should yield fairly accurate classification results. On the other hand, a lot of surnames might remain unclassified. For example, in Table 2, *Eisenhower* and *Khrushchell* are unclassified, because they do not contain strings that appear exclusively and frequently in one ethnic-name dictionary. However, it is still arguably true that they are of German and Russian ethnicity, respectively, given that some strings are largely (although not solely) associated with one language (e.g., *-Eis* in German, and *-Khr* in Russian); specifically, it is likely that *Eisenhower* is Americanized from the German name *Eisenhauer*, and *Khrushchell* is the misspelled version of *Khrushcheff*, which is Americanized from the Russian name *Khrushchev*. In this stage, I employ a probabilistic artificial intelligence algorithm—the naïve Bayes classifier (Rich, 2005)—to formalize the above idea. Suppose that there are n_i strings in surname i (denoted as $L_1, L_2, \cdots, L_{n_i}$), whose ethnicity is to be classified, and there are m possible categories of ethnicity (denoted as $e_1, e_2, \cdots, e_m$). For surname i, I calculate a "score" s^i_j for each category of ethnicity j (where $j = 1, 2, \cdots, m$):

$$s^i_j = \sum_{k=1}^{n_i} P(L_k) P(e^k_j | L_k) \tag{1}$$

where $P(L_k)$ is the frequency that the string L_k appears in training data (i.e., more common strings get higher "weights"), and $P(e^k_j | L_k)$ is the probability that L_k appears in ethnicity j's surname dictionary. Thus, i's ethnicity is $e = \arg\max s^i_j$, and in this stage, i's ethnicity remains unclassified only if there are ties. Essentially, this idea is an extended version of the deterministic algorithm used in the sub-stage 4 of the first stage, but allows ethnicity to be probabilistically determined. Table 3 shows that *Eisenhower* and *Khrushchell* are probabilistically classified as of German and Russian ethnicity.

Table 3: Second-Stage Classification

Surname	Place of birth	Mother tongue	Result	Stage
Smith	Poland	German	Anglicized	First stage, sub-stage 1
Dvorak	Germany	German	Others (Czech)	First stage, sub-stage 1
Bronstein	Poland	Yiddish	Jewish (by language)	First stage, sub-stage 2
Rabinovich	Russia	Russian	Jewish (by name)	First stage, sub-stage 2
Lisowski	Poland	Polish	Polish	First stage, sub-stage 3
Lvanov	Russia	Russian	Russian	First stage, sub-stage 4
Eisenhower	Russia	German	German	Second stage
Khrushchell	Russia	Russian	Russian	Second stage

3.1.3 Statistical Issues

I conclude the discussion of the classification algorithm by pointing out potential statistical issues. One major concern is that Americanized surnames are matched in the dictionary of Anglicized names, and are thus labeled as “unclassified”, but name Americanization is selected in terms of individual characteristics. For example, prior research finds that immigrants who localize their names usually have better labor market and social outcomes (e.g., Arai and Thoursie, 2009; Abramitzky et al., 2016; Xu, 2017). In Section 4, I will show that only 5% of surnames in actual census data are unclassified, which is consistent with earlier findings that surname Americanization was relatively rare (e.g., Biavaschi et al., 2017). Hence, selection should not threat the empirical analysis of this paper.

Another concern is measurement error. Using both the probabilistic and deterministic algorithm could increase the classification rate, but probabilistically classified ethnicity might be more likely to be misclassified. One way to investigate measurement errors is to redo the analysis based only on names classified deterministically. I will analyze potential impacts of measurement errors in the estimation in Section 5.3.

Finally, I am only able to classify ethnicity in the male census sample. This is because women usually changed surnames after marriage, and thus a female surname could not be used to infer ethnicity. Hence, the empirical conclusion of this paper—which will be based only on the male census sample—should be interpreted with caution, in the sense that male and female immigrants might have different settlement patterns.

3.2 The Performance in the Validation Dataset

Before employing the classification algorithm in actual census data, I first test its performance in a validation dataset, in which individuals’ information about ethnicity are known. I collect 750 German-ethnic, Polish-ethnic, and Russian-ethnic politicians who were born between 16th and 20th century; each ethnic group contains 250 names.[2] Similar to the historical context, in this validation dataset, an individual’s country of birth can be different from his ethnicity, e.g., Dmitry Foelkersam (who was a German-ethnic military politician in Russia). For this specific case, Foelkersam’s ethnicity is considered to be German, al-

[2] I do not show the list of names and reference due to space limitation, but they are available upon request.

though he was born in Russia, spoke Russian, and worked for the Russian Empire.

In artificial intelligence, two most important measures of the classification performance are *precision* and *recall* (Powers, 2011). Another relevant measure is the F-measure. For a specific category, they are calculated based on the following equations:

$$Precision_j = \frac{tp}{tp + fp}, Recall_j = \frac{tp}{tp + fn}, F_j = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2}$$

where for the ethnicity category j, tp represents the number of true positive cases that individuals in category j are correctly classified, fp represents the number of false positive cases that individuals in other categories are misclassified as in j, fn represents the number of false negative cases that individuals in j are misclassified as in other categories. Finally, the accuracy rate is a measure of overall classification results, which is the proportion of cases that individuals are correctly classified into the actual category.

Table 4: Classification Results in the Validation Dataset

	Precision	Recall	F-measure	Accuracy
German	0.95	0.98	0.96	
	(260)	(250)		
Polish	0.98	0.95	0.96	
	(242)	(250)		
Russian	0.95	0.94	0.95	
	(248)	(250)		
All origins				0.96
				(750)

Observations are in parentheses. Precision and recall are calculated based on different denominators (hence, difference values of observations).

Table 4 presents classification results. The first column shows that among 260 individuals that are classified as of German ethnicity, 95% of them were indeed German ethnics; among 250 German ethnics in this validation dataset, the classification algorithm finds 98% of them. The F-measure is 0.96, which is very close to 1. I observe similar classification results for Polish and Russian ethnics in the validation dataset, and the overall accuracy rate of classification is 96%. The above results suggest that this artificial intelligence algorithm is arguably reliable for surname-based ethnicity classification.

4 Data and Methods

In this section, I first introduce the 1920 and 1930 U.S. census immigrant sample. I then report the descriptive statistics. Subsequently, I discuss empirical models for studying within-group ethnic segregation.

4.1 Data

In this paper, I use the 1920 and 1930 full-count U.S. census (Ruggles et al., 2017) to generate the sample, which contains all first-generation male immigrants born in Germany, Poland, or Russia. I present statistics of three types of origins in Table 5. In each panel—showing the 1920 and 1930 census, respectively—I first list two traditional types of origins: the country of birth and mother tongue. In 1920, 37.4%, 27.8%, and 34.9% of all individuals in this sample were born in Germany, Poland, and Russia. 40.1% and 24.5% of immigrants in the sample spoke German and Polish, respectively; these numbers appear to be close to the proportion of German-born and Polish-born immigrants. However, only 9.3% of immigrants spoke Russian. In other words, many Russian-born immigrants did not speak Russian, which could be due to the large non-Russian-ethnic immigrant population from Russia (Haines, 2000). 20.6% of immigrants spoke a Jewish language (i.e., Yiddish or Hebrew) as the mother tongue.

Table 5: Basic Demographic Characteristics: Origin

	German	Polish	Russian	Jewish	Others
A. 1920:					
Country of birth	0.374	0.278	0.349	—	—
Mother tongue	0.401	0.245	0.093	0.206	0.055
Name-based ethnicity	0.489	0.144	0.096	0.224	0.048
B. 1930:					
Country of birth	0.391	0.314	0.296	—	—
Mother tongue	0.419	0.238	0.078	0.233	0.032
Name-based ethnicity	0.484	0.142	0.083	0.248	0.043

Observations: 2,309,167 (1920); 2,154,493 (1930).

The differences in the first two measures of origin suggest possible heterogeneity in ethnicity in the sample. Indeed, the third row in Panel A shows that 48.9% of immigrants in the sample had German surnames; 14.4% of immigrants had Polish surnames; finally, 9.6% of immigrants had Russian surnames. Note that most cases of Jewish ethnicity in the sample are identified based on the mother tongue (20.6% versus 22.4%). The disproportionately large population associated with German surnames could be due to ethnic German settlements in Eastern Europe (Sammartino, 2010). Some Jews also adopted German surnames following laws in the late eighteenth century (Dubin, 1999), although most of them are still identified based on the mother tongue.

I observe similar patterns of the country-of-birth, mother tongue, and surname-based origin in the 1930 sample in Panel B: three types of origins are not always consistent. Compared with Panel A, the major difference is that there were relatively fewer Russian-born and Russian-speaking immigrants, and more Jews in 1930. This is not surprising, as prior research points out that Russian Jews had more incentives to move to the U.S.; furthermore, they were less likely to return back to Europe (Boustan, 2007; Greenwood and Ward, 2015; Ward, 2017), although return migration was common among European immigrants in the early twentieth century (Haines, 2000; Abramitzky et al., 2014).

Table 6: The Geography (Birthplace) of Surname-Based Ethnicity

Surname-based ethnicity:	German	Polish	Russian	Jewish	Others	Total
A. 1920:						
Germany	0.832	0.028	0.052	0.015	0.072	1
Poland	0.300	0.411	0.151	0.102	0.037	1
Russia	0.272	0.055	0.098	0.545	0.031	1
B. 1930:						
Germany	0.846	0.027	0.049	0.015	0.064	1
Poland	0.258	0.391	0.126	0.199	0.034	1
Russia	0.245	0.031	0.082	0.616	0.026	1

Observations: 2,309,167 (1920); 2,154,493 (1930).

Table 6 summarizes the geographic distribution of surname-based ethnicity by the country of birth. Panel A studies the 1920 sample. In 1920, 83.2% of all German-born immigrants in the sample were associated with German surnames. There were few German-born immigrants associated with other origins. Among German-born immigrants, 7.2% of surnames belonged to origins other than German, Polish, Russian, and Jewish; most of them had Anglicized names, which is not surprising because German immigrants were culturally more assimilated (Abramitzky et al., 2016).[3] 41.1% of Polish-born immigrants were associated with Polish surnames. Although Polish surnames were most common in the Polish population, there were still many Polish-born immigrants who had German and Russian surnames; also, 10.2% of Polish immigrants in the sample were Jews. Finally, I find that only 9.8% of Russian-born immigrants in the sample were associated with Russian surnames. On the other hand, nearly 30% of Russian immigrants had German surnames, and more than half of Russian immigrants were Jews. Given the political situation in the new Soviet republic (Boustan, 2007), it is not surprising that Russian immigration to the U.S. was highly selected in terms of demographic characteristics, and ethnic minorities (e.g., German ethnics and Jews) had more incentives to leave Russia. Russian ethnics might also have less information about U.S. immigration in the early twentieth century (Haines, 2000). Panel B focuses on the 1930 sample. The major differences between two panels is that there was a huge increase in the Eastern European Jewish population. As discussed earlier, a key reason was the low return migration rate among Jews (Ward, 2017).

In Table 7 and 8 I present descriptive statistics of the 1920 sample by three types of origins. In Table 7 I focus on basic demographic and socioeconomic variables. The first panel shows descriptive statistics by the country of birth. In general, German-born immigrants were older and stayed in the U.S. longer, but three groups had the similar average age of immigration (approximately 20 years old). German-born and Polish-born immigrants were more likely to be married than Russian-born immigrants. German-born immigrants had a significant higher citizenship rate, lower rate of urban residence, higher rate of farm residence, and higher rate of homeownership. German-born immigrants had lowest average

[3] Another possible reason behind the large population of Anglicized names among German immigrants in 1920 is that Germany was in the Central Powers against the U.S. in the World War I, and many German immigrants in the U.S. converted to Anglicized surnames to avoid hostility towards them.

Table 7: Descriptive Statistics by Origin (1): Basic Variables, 1920

	Age	Year since	Married	Citizen-	Urban	Farm	Occupa-	Homeowner-	Obser-
		migration		ship	status	status	tional score	ship	vations
German-born	52.003	32.917	0.722	0.766	0.662	0.187	15.496	0.573	862,925
	(15.553)	(15.148)	(0.448)	(0.423)	(0.473)	(0.390)	(13.738)	(0.494)	
Polish-born	37.641	16.550	0.723	0.298	0.823	0.058	17.777	0.353	641,344
	(13.374)	(10.659)	(0.447)	(0.458)	(0.381)	(0.234)	(11.875)	(0.478)	
Russian-born	35.767	15.881	0.678	0.410	0.876	0.057	18.853	0.272	804,898
	(13.371)	(9.081)	(0.467)	(0.492)	(0.330)	(0.231)	(14.895)	(0.445)	
German-speaking	51.099	31.998	0.724	0.748	0.637	0.209	15.277	0.581	940,071
	(15.962)	(15.280)	(0.447)	(0.434)	(0.481)	(0.407)	(13.561)	(0.493)	
Polish-speaking	36.824	15.616	0.719	0.263	0.829	0.048	17.762	0.348	566,515
	(12.688)	(9.798)	(0.449)	(0.440)	(0.376)	(0.214)	(11.366)	(0.476)	
Russian-speaking	35.320	14.888	0.657	0.369	0.858	0.041	18.942	0.235	214,166
	(12.403)	(9.128)	(0.475)	(0.483)	(0.349)	(0.198)	(14.617)	(0.424)	
German-name†	46.906	27.300	0.714	0.628	0.692	0.165	16.167	0.500	1,129,048
	(16.468)	(16.431)	(0.452)	(0.483)	(0.462)	(0.371)	(13.521)	(0.500)	
Polish-name†	37.479	16.432	0.718	0.294	0.835	0.055	17.499	0.384	331,657
	(13.213)	(10.544)	(0.450)	(0.455)	(0.371)	(0.228)	(11.461)	(0.486)	
Russian-name†	38.759	17.672	0.680	0.344	0.771	0.078	17.478	0.334	220,522
	(14.321)	(12.611)	(0.466)	(0.475)	(0.420)	(0.228)	(12.311)	(0.472)	
Jewish	36.341	16.866	0.700	0.453	0.966	0.011	19.799	0.238	516,909
	(14.095)	(9.386)	(0.458)	(0.498)	(0.180)	(0.104)	(15.775)	(0.426)	

Standard deviations are in parentheses. †: by definition, the Jews are *not* included in the third panel (name-based ethnicity).

occupational scores (conditional on employment, however, this pattern no longer exists). In the second panel I show statistics by the mother tongue, and the pattern appears to be very similar to that reported in the first panel.

The third panel, which presents statistics by surname-based ethnicity, suggests that surname-based ethnicity could be significantly different from the first two types of origins. Compared with immigrants born in Germany or speaking German, immigrants of German surname-based ethnicity were younger, less likely to be citizens, more likely to live in cities, less likely to live in farms, and less likely to own a house. This panel also shows some differences among three types of Polish (and Russian) origins. The last panel focuses on the Jewish population, identified based on the mother tongue and name. Most Jewish immigrants lived in cities, and had relatively high occupational scores.

In Table 8 I focus on settlement patterns by three types of immigrant origins. I report the number of immigrants of specific origins at the level of the enumeration district (ED), which was the smallest geographic unit in the 1920 and 1930 census, and usually contained less than 2,000 residents. I list both the full and non-Jewish population of Eastern European

Table 8: Descriptive Statistics by Origin (2): Settlement Patterns, 1920

	The number of immigrants in the local enumeration district (ED), by origin										
	German origin		Polish origin			Russian origin			Jews	Native	Obser-
	Birth	Language	Birth†	Birth‡	Language	Birth†	Birth‡	Language			vations
German-born	72	96	29	27	28	29	15	10	20	1,214	862,925
	(72)	(100)	(108)	(106)	(116)	(84)	(46)	(37)	(87)	(641)	
Polish-born	43	58	371	354	366	88	47	38	70	1,396	641,344
	(51)	(74)	(508)	(510)	(521)	(179)	(118)	(87)	(206)	(761)	
Russian-born	34	65	74	53	67	314	92	64	291	1,171	804,898
	(46)	(102)	(197)	(190)	(204)	(368)	(182)	(149)	(423)	(684)	
German-speaking	70	107	28	26	24	38	25	10	18	1,203	940,071
	(71)	(117)	(98)	(96)	(99)	(105)	(81)	(39)	(82)	(637)	
Polish-speaking	44	54	399	394	422	80	55	37	33	1,438	566,515
	(52)	(67)	(526)	(526)	(537)	(162)	(132)	(87)	(117)	(772)	
Russian-speaking	37	66	91	83	92	217	157	153	87	1,246	214,166
	(47)	(80)	(247)	(245)	(249)	(307)	(238)	(233)	(213)	(780)	
German-name	62	94	80	77	80	68	47	32	30	1,237	1,129,048
	(67)	(109)	(248)	(246)	(255)	(169)	(134)	(106)	(116)	(679)	
Polish-name	49	59	420	417	442	74	56	37	26	1,453	331,657
	(55)	(72)	(551)	(550)	(562)	(157)	(135)	(84)	(95)	(782)	
Russian-name	42	64	193	188	200	109	76	59	47	1,322	220,552
	(54)	(86)	(387)	(386)	(395)	(214)	(168)	(133)	(153)	(751)	
Jewish	32	47	78	33	40	389	44	37	464	1,114	516,909
	(44)	(68)	(137)	(98)	(105)	(390)	(101)	(96)	(463)	(618)	

Standard deviations are in parentheses. †: full population, including Jews. ‡: non-Jewish population.

origin, i.e., all Polish/Russian immigrants in the ED, and non-Jewish Polish/Russian immigrants in the ED. The first panel shows evidence of "country-of-birth enclave residence", i.e., immigrants lived in the ED with much more immigrants born in the same country. For example, on average, German-born immigrants lived in EDs with 72 immigrants born in Germany, a number higher than that for Polish-born and Russian-born immigrants (43 and 34). Similarly, the second panel presents evidence of language enclave residence.

The first two panels show that the country-of-birth origin and mother-tongue origin was similar for immigrants in the 1920 sample. However, surname-based ethnicity could be very different from the first two types of origins. While the differences among three types of German origins were relatively small in 1920, there were huge differences among three types of Polish and Russian origins. Specifically, immigrants of Russian surname-based ethnicity lived in EDs with significantly *fewer* Russian-born immigrants, compared with all immigrants born in Russia (109 v.s. 314). I find similar results for the Polish population. These differences among Eastern European immigrants could be due to the large Jewish population. Indeed, there were much smaller differences in local non-Jewish

Russian-born and non-Jewish Russian-speaking population between immigrants of Russian surname-based ethnicity and Russian-born immigrants (76 v.s. 92; 59 v.s. 64). Similarly, immigrants of Russian surname-based ethnicity also had much fewer Jewish neighbors in the ED than Russian-born immigrants (47 v.s. 291).

Table 9: Descriptive Statistics by Origin (1): Basic Variables, 1930

	Age	Year since migration	Married	Citizen-ship	Urban status	Farm status	Occupa-tional score	Homeowner-ship	Obser-vations
German-born	51.197	32.307	0.670	0.734	0.714	0.152	16.061	0.576	841,516
	(17.698)	(19.040)	(0.470)	(0.442)	(0.452)	(0.359)	(14.021)	(0.494)	
Polish-born	43.195	22.042	0.769	0.563	0.854	0.057	18.891	0.505	675,352
	(12.481)	(9.983)	(0.422)	(0.496)	(0.353)	(0.023)	(12.776)	(0.500)	
Russian-born	42.283	23.063	0.777	0.683	0.894	0.055	20.870	0.363	637.625
	(13.083)	(9.990)	(0.416)	(0.465)	(0.308)	(0.229)	(15.749)	(0.481)	
German-speaking	50.798	31.901	0.681	0.728	0.695	0.170	16.050	0.580	902,555
	(17.532)	(18.662)	(0.466)	(0.445)	(0.461)	(0.375)	(13.878)	(0.493)	
Polish-speaking	43.838	23.708	0.773	0.533	0.830	0.066	18.469	0.559	509,919
	(11.918)	(9.751)	(0.419)	(0.499)	(0.375)	(0.248)	(12.002)	(0.497)	
Russian-speaking	41.922	21.939	0.736	0.598	0.863	0.046	20.678	0.363	167,306
	(12.040)	(9.731)	(0.441)	(0.490)	(0.344)	(0.209)	(15.379)	(0.481)	
German-name†	48.677	29.472	0.700	0.689	0.732	0.141	16.931	0.547	1,041,919
	(16.718)	(17.366)	(0.458)	(0.463)	(0.443)	(0.349)	(13.977)	(0.498)	
Polish-name†	44.450	24.420	0.779	0.542	0.835	0.070	18.184	0.595	306,164
	(12.233)	(10.462)	(0.415)	(0.498)	(0.371)	(0.256)	(11.889)	(0.491)	
Russian-name†	44.327	24.187	0.720	0.555	0.780	0.088	18.243	0.492	178,786
	(13.724)	(12.843)	(0.449)	(0.457)	(0.414)	(0.284)	(13.094)	(0.500)	
Jewish	42.008	22.699	0.781	0.725	0.975	0.009	21.546	0.314	534,693
	(13.803)	(10.372)	(0.414)	(0.446)	(0.156)	(0.092)	(16.150)	(0.464)	

Standard deviations are in parentheses. †: by definition, the Jews are *not* included in the third panel (name-based ethnicity).

I present descriptive statistics of the 1930 sample in Table 9 and 10. Table 9 shows that the basic demographic and socioeconomic characteristics among German-born (and German-speaking) immigrants in 1930 were similar to those in 1920. However, Table 9 presents differences in individual characteristics among Eastern European immigrants between 1920 and 1930. In particular, for Eastern European immigrants (classified by either the country of birth or mother tongue), the age, years since migration, citizenship rate, and occupational scores significantly increased in the 1920s. This was also true for Jewish immigrants. The above results could be related to the effects of immigration restriction laws in the early 1920s, which imposed severe limitations on new immigration from Eastern Europe, and much less severe limitations on German immigration. Similar to Table 7, Table 9 shows that compared with the country of birth and mother tongue, surname-based ethnicity

Table 10: Descriptive Statistics by Origin (2): Settlement Patterns, 1930

	The number of immigrants in the local enumeration district (ED), by origin										
	German origin		Polish origin			Russian origin			Jews	Native	Obser-
	Birth	Language	Birth†	Birth‡	Language	Birth†	Birth‡	Language			vations
German-born	71	93	26	22	21	28	10	10	24	1,311	841,516
	(81)	(109)	(67)	(62)	(61)	(80)	(30)	(28)	(92)	(776)	
Polish-born	34	47	206	172	166	106	21	31	116	1,375	675,352
	(44)	(63)	(213)	(209)	(210)	(212)	(46)	(60)	(271)	(694)	
Russian-born	35	61	86	36	34	263	48	43	293	1,301	637,625
	(47)	(88)	(118)	(73)	(71)	(296)	(111)	(97)	(381)	(741)	
German-speaking	69	96	25	21	20	32	16	10	23	1,294	902,555
	(80)	(116)	(65)	(59)	(58)	(93)	(59)	(33)	(89)	(788)	
Polish-speaking	35	47	216	(209)	207	44	19	27	33	1,408	509,919
	(45)	(62)	(223)	(222)	(223)	(110)	(45)	(56)	(120)	(700)	
Russian-speaking	38	60	83	64	58	148	75	84	99	1,334	167,306
	(48)	(76)	(126)	(110)	(102)	(232)	(151)	(144)	(203)	(763)	
German-name	61	86	53	47	44	49	24	22	35	1,311	1,041,919
	(75)	(108)	(118)	(113)	(111)	(129)	(78)	(67)	(120)	(775)	
Polish-name	38	50	224	219	216	31	16	25	20	1,417	306,164
	(49)	(67)	(226)	(225)	(227)	(78)	(35)	(49)	(85)	(700)	
Russian-name	40	58	118	109	102	66	32	41	47	1,362	178,786
	(55)	(80)	(174)	(170)	(168)	(147)	(83)	(89)	(141)	(746)	
Jewish	34	51	121	31	30	346	29	32	431	1,304	534,693
	(44)	(64)	(142)	(59)	(57)	(311)	(59)	(63)	(411)	(687)	

Standard deviations are in parentheses. †: full population, including Jews. ‡: non-Jewish population.

was a different type of origin in the 1930 U.S.

Table 10 presents settlement patterns by origin in the 1930 sample. In general, the descriptive findings of Table 10 are similar to those reported in the 1920 sample in Table 8: I observe clear evidence of country-of-birth and language enclave residence for all groups. Again, I find that compared with the country of birth and mother tongue, surname-based ethnicity led to significantly different settlement patterns in the 1930 sample. In sum, Table 7 to 10 indicate that three different types of immigrant origins lead to different descriptive results. In particular, an immigrant's surname-based ethnicity was not necessarily consistent with two other types of origins in both the 1920 and 1930 sample.

4.2 Empirical Strategies

I now discuss the empirical strategies of this paper. I first aggregate individual records in the 1920 and 1930 sample to the ED and county level, and calculate ethnic segregation within an immigrant group (defined based on the country of birth) by the widely used dissimilarity index of segregation (Duncan and Lieberson, 1959; Winship, 1977):

$$D_{ab}^{jk} = \frac{1}{2} \sum_{t=1}^{n_j} |\frac{a_t}{a_j} - \frac{b_t}{b_j}| \tag{3}$$

where j indexes the county, and k indexes the country of birth. Within this immigrant group, a and b denote two sub-groups that have different surname-based ethnicity. n_j is the total number of EDs in county j. a_j is the total number of k-born immigrants with a-origin surnames in county j, and a_t is the number of such immigrants in the t-th ED. I similarly define b_j and b_t. D_{ab}^{jk} is the degree of within-group (k) dissimilarity between a and b in county j. By this definition, the degree of dissimilarity should be small if the ethnic differences in settlement patterns within an immigrant group from the same country of birth were small. On the other hand, a high degree of dissimilarity suggests significant ethnic heterogeneity in settlement patterns.

The above measure studies county-level ethnic segregation within each immigrant group. There are, however, at least two reasons to further study ethnic segregation based on individual records. First, one can include covariates to account for the effects of individual characteristics. Second, while the dissimilarity index is measured at the county level, one can further explore the "micro-structure" of ethnic segregation by focusing on the ED (i.e., sub-county) level within a regression framework based on individual records. I estimate the OLS specification *within* each immigrant group defined based on the country of birth:

$$N_{iek} = \alpha + \mathbf{E}_{ik}\beta + \gamma J_i + \delta \hat{N}_e + \mathbf{X}_i\mu + \tau_{i(s)} + \varepsilon_{iek} \tag{4}$$

where i indexes the immigrant, e indexes the ED where i lived, and k indexes i's country of birth. $\mathbf{E}_{ik}$ is the vector of key variables of interest, which are indicators of surname-based ethnicity (within the immigrant group born in country k), and the majority group (whose ethnicity is consistent with the country of birth, e.g., Polish ethnics are considered as majorities among Polish-born immigrants) is omitted in $\mathbf{E}_{ik}$. Note that the algorithm mainly focuses on classifying the German, Polish, and Russian origin; I further include an indicator of Jewish ethnicity J_i. $\hat{N}_e$ is the total number of residents (including natives) in ED e, $\mathbf{X}_i$ is the vector of individual characteristics (e.g., age, years since migration, urban status), and $\tau_{i(s)}$ are state fixed effects. I cluster the standard errors at the state level.

Within an immigrant group from the same country of birth, the coefficients β should be insignificant if different ethnic groups were not spatially isolated from each other. On the other hand, the significant coefficients β suggest within-group ethnic heterogeneity in settlement patterns. Without a valid instrument in cross-section data, Equation 4 is unlikely to reveal the causal relationship between ethnicity and settlement patterns. However, one can establish additional specifications to study various factors that could potentially affect the causal interpretation of the empirical results.

I first explore measurement errors. In Equation 4, $\mathbf{E}_{ik}$ contains binary indicators of surname-based ethnicity. The estimates might be biased due to measurement errors if individuals' ethnic origins are misclassified by the algorithm. I consider an additional test based on the sub-sample of "more reliable cases of ethnicity": recall that the first-stage classification algorithm is a deterministic algorithm, and arguably yields low degrees of measurement errors, as surnames need to be matched perfectly or based on linguistic elements that are exclusively used in surnames in specific languages. Hence, I focus on the sub-sample that contains individuals whose surnames are classified in the first stage, and examine whether results of this test are consistent with main results.

I then study effects of other geographic characteristics on segregation. Economists find that immigrants are more likely to reside in areas with previous immigrants of the same origin (Bartel, 1989; Altonji and Card, 1994). Does ethnicity only coincidently relate to ethnic enclave residence and ethnic segregation? The major threat is that immigrants of the same origin were concentrated not because they chose ethnic enclaves, but because there were some geographic characteristics that attracted co-ethnics to reside together, i.e., an area has some geographic characteristics (such as climatic patterns, soil characteristics, and occupational agglomeration) that particularly attracted a specific ethnic group. To study this, I propose three additional sets of tests. First, I include county fixed effects to capture the fact that some counties are particularly "gateway counties" for some ethnic groups. Second, I estimate heterogeneous effects of ethnicity in both the urban and rural sub-sample, and see if results in the urban and rural sub-sample are similar. Finally, I study whether ethnic segregation was driven by local labor market characteristics. In general, it is possible that immigrants choose ethnic enclave residence because ethnic networks provide employment opportunities (Edin et al., 2003; Munshi, 2003), and thus ethnicity has a causal effect on settlement patterns. However, if "suitable jobs" for immigrants are spatially concentrated, then segregation could simply be due to occupational agglomeration. To examine this, I reestimate Equation 4 by including ED-level labor market characteristics.

5 Results

This section reports the empirical results of this paper. I first present results of within-group ethnic segregation at the county level. I then explore the micro-structure of segregation at the ED level based on individual census records within a regression framework. Finally, I discuss several statistical issues of the empirical strategies, including measurement errors and other factors that are likely to be correlated with both ethnicity and segregation.

5.1 Ethnic Segregation at the County Level

Table 11 presents county-level ethnic segregation within each immigrant group from the same country of birth. Panel A shows the segregation pattern in 1920. I first focus on German-born immigrants, and study the dissimilarity index of segregation (introduced in Section 4.2) between immigrants of German surname-based ethnicity and immigrants of other ethnicity. I find that the degree of ethnic segregation was very high among German-born immigrants: German ethnics (i.e., those with German surnames) were spatially isolated from their compatriots with other surname-based ethnicity, as well as Jews. I present the degrees of segregation weighted by the county population in brackets, and find very similar results. Subsequently, I study Polish-born and Russian-born immigrants. Results show that the degrees of segregation between Polish ethnics and other ethnics born in Poland were around or above 0.5, although the numbers were smaller than those in the German-born immigrant group. Among Polish-born immigrants, Polish Jews were again more likely to live in segregated areas. I find similar results among Russian-born immigrants, where I define either immigrants of Russian surname-based ethnicity (the majority group back in Russia) or immigrants of German surname-based ethnicity (the majority group among Russian immigrants in the U.S.) as ethnic majorities. I repeat the exercise in the 1930 sample in Panel B and observe similar segregation patterns, i.e., the degree of ethnic segregation within each immigrant group was generally high.

Table 11: Ethnic Segregation within the Immigrant Group

Surname-based ethnicity	German	Polish	Russian	Jewish
A. 1920:				
German-born, German-name	—	0.70	0.60	0.73
	—	[0.72]	[0.55]	[0.76]
Polish-born, Polish-name	0.52	—	0.49	0.72
	[0.48]	—	[0.46]	[0.87]
Russian-born, Russian-name	0.52	0.54	—	0.74
	[0.45]	[0.55]	—	[0.79]
Russian-born, German-name	—	0.57	0.52	0.72
	—	[0.59]	[0.45]	[0.76]
B. 1930:				
German-born, German-name	—	0.73	0.66	0.72
	—	[0.72]	[0.58]	[0.74]
Polish-born, Polish-name	0.53	—	0.51	0.72
	[0.50]	—	[0.49]	[0.88]
Russian-born, Russian-name	0.58	0.60	—	0.72
	[0.52]	[0.64]	—	[0.77]
Russian-born, German-name	—	0.62	0.58	0.71
	—	[0.67]	[0.52]	[0.73]

The degrees of segregation weighted by the county population are shown in brackets.

How high (or low) are the degrees of segregation shown in this table? I compare these numbers with two sets of segregation measures. First, Table 12 shows the degree of im-

Table 12: Immigrant Segregation with the Native-Born Population

Country of birth:	Germany	Poland	Russia	Yiddish/Hebrew
1920 census	0.42	0.74	0.66	0.79
1930 census	0.47	0.76	0.70	0.81

Table 13: Segregation within Each Surname Group

Surname-based ethnicity	German	Polish	Russian	Jewish
1920 census	0.60	0.65	0.69	0.57
1930 census	0.64	0.66	0.69	0.54

The degrees of segregation weighted by the county population are shown in brackets.

migrant segregation with the native-born population. Results of two tables suggest that the degree of German-native segregation was significantly lower than the degree of ethnic segregation within the German-born immigrant group, although ethnic segregation within the Polish-born (and Russian-born) immigrant group appeared to be relatively lower than Polish-native (and Russian-native) segregation.

Second, in Table 13 I study segregation within each surname origin group. I examine segregation within the German surname group by calculating the degree of segregation between German-born German ethnics and immigrants with German surnames born in Poland or Russia. I similarly calculate segregation for the Polish and Russian surname group. I also calculate segregation between Polish and Russian Jews (over 90% of Yiddish or Hebrew speakers were born in Poland or Russia). Table 13 indicates that segregation by birthplace within the German surname group was actually lower than ethnic segregation among German-born immigrants. Similarly, segregation within Jews was relatively low. Segregation within the Polish (and Russian) surname group was higher than ethnic segregation among Polish-born (and Russian-born) immigrants.

The above tables suggest that ethnic differences in settlement patterns within each immigrant group did exist in 1920 and 1930. In particular, immigrants of German and Jewish ethnicity lived more closely with immigrants of same ethnicity, rather than other immigrants born in the same country. Hence, surname-based ethnicity could provide important demographic information in addition to the country of birth.

5.2 Main Results: Country-of-Birth Enclave Residence

The above county-level analysis presents descriptive results of within-group ethnic segregation. I now further investigate the relationship between ethnicity and immigrant enclave residence based on individual records within a regression framework. By doing so, I am able to control for individual characteristics, such as age, years since migration, and urban status. I am also able to explore ethnic enclave residence at the ED level.

The classical conclusion of immigrant enclave residence is that immigrants prefer to reside in areas with more immigrants born in the same country of origin (e.g., Bartel, 1989; Altonji and Card, 1994), i.e., "country-of-birth enclave residence". To estimate the relationship between ethnicity and country-of-birth enclave residence, I focus on each subsample of immigrants by the country of birth, and regress the number of immigrants born in a specific country at the ED level on surname-based ethnicity, following Equation 4.

Table 14: Main Results: Within-Group Ethnic Segregation, 1920

	The Number of Immigrants Born in the Same Country of Origin (The Size of the Country-of-Birth Enclave), ED								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample:	Germany	Poland	Poland†	Poland‡	Poland†‡	Russia	Russia†	Russia‡	Russia†‡
Average:	72.701	372.897	393.965	355.432	389.090	318.203	192.323	90.147	147.893
German-name		−76.982***	−79.305***	−74.068***	−81.839	1.038	16.708***	46.491***	18.291***
		(12.910)	(12.185)	(13.330)	(13.437)	(8.180)	(2.933)	(14.820)	(3.788)
Polish-name	−0.543					−62.411**	−28.103	10.759	−9.956
	(2.953)					(19.443)	(17.013)	(10.042)	(14.561)
Russian-name	−4.995***	−64.422***	−61.352***	−64.422***	−69.062***				
	(0.785)	(9.581)	(9.197)	(9.745)	(9.284)				
Jewish	−9.971**	−125.835***		−233.334***		174.958***		−83.393***	
	(3.283)	(22.680)		(20.141)		(15.018)		(13.777)	
Yrs. since migration	−0.189***	−0.739*	−0.613*	−0.627*	−0.570*	−1.653***	−0.886**	−0.439**	−0.650*
	(0.041)	(0.288)	(0.261)	(0.249)	(0.258)	(0.336)	(0.246)	(0.145)	(0.266)
Total ED residents	0.034***	0.264***	0.272***	0.263***	0.272***	0.172***	0.109***	0.062***	0.085***
	(0.004)	(0.043)	(0.042)	(0.043)	(0.042)	(0.021)	(0.019)	(0.013)	(0.019)
Adj. R^2	0.322	0.638	0.654	0.645	0.653	0.436	0.309	0.261	0.263
Obs.	800,499	617,538	552,234	617,538	552,234	780,129	341,747	780,129	341,747

Standard errors are clustered at the state level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *state* fixed effects.
†: Only non-Jewish immigrants born in the specific country of origin are included in the sample (Jewish dummy is omitted in the model).
‡: Only non-Jewish immigrants born in the specific country of origin are included in the dependent variable (Jews are not considered in enclaves).

Table 14 presents main results in the 1920 sample. I exclude immigrants with Anglicized surnames and surnames of other European origins, but keep immigrants classified as Jews. In Column 1 I focus on German-born immigrants, and regress the number of

German-born immigrants at the ED level on ethnicity indicators and other individual characteristics. I control for state fixed effects, and cluster standard errors at the state level. Results show that compared with those of German surname-based ethnicity, German-born immigrants with Polish surnames had similar patterns of German enclave residence. However, those of Russian surname-based ethnicity and German Jews lived in EDs with fewer German-born immigrants.

I then turn to study the sample of the Polish-born immigrant group. In Column 2, I find that compared with those of Polish surname-based ethnicity, immigrants from all other surname-based groups resided in areas with much fewer Polish-born immigrants. I exclude Polish Jews in the sample in Column 3, and find almost the identical results. In Column 4 I redefine Polish enclaves by focusing only on non-Jewish Polish immigrants at the ED level. I find that (a) the magnitudes of the coefficients for the German-name and Russian-name indicator remain unchanged, and (b) the magnitude of the coefficient for Polish Jews becomes much larger, which is unsurprising, as Jews were more isolated from non-Jews. In Column 5 I exclude Polish Jews from both the sample and the dependent variable. Results are quantitatively similar to those reported in earlier columns.

Table 15: Main Results: Within-Group Ethnic Segregation, 1930

	The Number of Immigrants Born in the Same Country of Origin (The Size of the Country-of-Birth Enclave), ED								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample:	Germany	Poland	Poland†	Poland‡	Poland†‡	Russia	Russia†	Russia‡	Russia†‡
Average:	71.902	207.048	212.127	172.042	204.442	238.167	118.438	47.147	76.237
German-name		−55.201***	−53.584***	−53.941***	−59.544***	5.195	18.237***	28.137***	15.943**
		(7.238)	(6.083)	(7.530)	(8.869)	(6.891)	(3.918)	(3.935)	(4.478)
Polish-name	−6.883***					−44.857***	−24.630***	−6.935*	−13.131***
	(1.835)					(8.999)	(5.497)	(2.436)	(3.293)
Russian-name	−5.521***	−39.149***	−38.288***	−37.146***	−42.355***				
	(0.892)	(4.746)	(4.018)	(4.716)	(5.358)				
Jewish	−10.058**	−66.629*		−180.903		132.984***		−38.858***	
	(3.480)	(30.384)		(19.530)		(16.181)		(4.304)	
Yrs. since migration	−0.384***	−0.530*	−0.320	−0.088	−0.154	−2.007***	−0.657**	−0.215**	−0.373
	(0.080)	(0.197)	(0.270)	(0.200)	(0.300)	(0.350)	(0.224)	(0.093)	(0.202)
Total ED residents	0.034***	0.093***	0.099**	0.079**	0.096**	0.126***	0.087***	0.033**	0.061**
	(0.006)	(0.023)	(0.027)	(0.026)	(0.027)	(0.015)	(0.021)	(0.010)	(0.020)
Adj. R^2	0.344	0.347	0.380	0.390	0.370	0.422	0.313	0.194	0.246
Obs.	787,531	652,810	523,251	652,810	523,251	621,221	228,270	621,221	228,270

Standard errors are clustered at the state level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *state* fixed effects.
†: Only non-Jewish immigrants born in the specific country of origin are included in the sample (Jewish dummy is omitted in the model).
‡: Only non-Jewish immigrants born in the specific country of origin are included in the dependent variable (Jews are not considered in enclaves).

From Column 5 to 9 I repeat the exercise in the 1920 sample of Russian-born immigrants. Among Russian-born immigrants, 54.5% were Jews, 27.2% were of German surname-based ethnicity, and only 9.8% of Russians were of Russian surname-based ethnicity. This could be because that German ethnics and Jews were more likely to be affected by political instability in Russia before and around 1920, and had more incentives to migrate. Column 6 shows that Russian Jews lived in significantly larger Russian enclaves, mainly because of the large Russian Jewish population in the U.S. I exclude Russian Jews in the sample and rerun the regression in Column 7. Results show that compared with those of Russian surname-based ethnicity—which were ethnic majorities back in Russia—Russian immigrants of German surname-based ethnicity lived in larger Russian enclaves. Again, this could be because German ethnics were majorities among non-Jewish Russians in the U.S. Hence, although surprising at first glance, Table 14 does present reasonable results that reflect the demographics of Russian immigrants in 1920.

In Table 15, I redo the above empirical analysis in the 1930 sample. I find generally similar patterns of within-group ethnic segregation. Compared with Table 14, the magnitudes of the coefficients in most specifications become smaller, but still appear to be statistically significant. In general, immigrant enclaves declined during the 1920s, along with the process of urbanization associated with internal migration. This was especially true for Southern and Eastern European immigrant enclaves, as immigration restriction laws in the early 1920s severely limited immigration from Italy, Poland, and Russia.

An interesting question related to Russian immigration is: were Russian immigrants who had German surnames Jews, or actually German ethnics? Following laws regarding cultural assimilation, German surnames were common among Jews (Dubin, 1999); on the other hand, according to the 1897 Russian census, most Russian Jews spoke Yiddish rather than other languages (Kreindler, 1985). Table 14 and 15 show that although some Russians with German surnames might be Jews even if they did not spoke Yiddish or Hebrew, most of such individuals should be of German ethnicity: Column 8 of both tables show that Russian

Jews and Russian immigrants with German surnames were highly isolated from each other. Given the low degree of segregation within the Jewish population, these immigrants with German surnames were probably indeed German ethnics, but not Jews.

Table 16: Within-Group Ethnic Segregation, Population Shares

	The Percentage (%) of Immigrants Born in the Same Country of Origin (The Size of the Country-of-Birth Enclave), ED									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	1920 Sample, (1) - (5)					1930 Sample, (6) - (10)				
Sample:	Germany	Poland	Poland†‡	Russia	Russia†‡	Germany	Poland	Poland†‡	Russia	Russia†‡
Average:	4.845	14.590	14.808	16.422	7.827	4.518	10.453	10.281	11.608	4.573
German-		−3.642***	−3.897***	−0.297	0.833***		−2.840***	−3.034***	0.151	0.638***
name		(0.502)	(0.495)	(7.238)	(0.200)		(0.399)	(0.471)	(0.296)	(0.121)
Polish-	−0.075			−2.808**	−0.624	−0.400***			−2.280***	−0.731***
name	(0.140)			(0.907)	(0.626)	(0.105)			(0.441)	(0.153)
Russian-	−0.314***	−2.737***	−2.980***			−0.325***	−2.000***	−2.157***		
name	(0.049)	(0.368)	(0.360)			(0.046)	(0.279)	(0.310)		
Jewish	−0.616**	−5.056*		−8.702***		−0.550**	−3.058*		6.443***	
	(0.183)	(1.945)		(30.384)		(0.199)	(1.464)		(0.696)	
Yrs. since	−0.014***	−0.051***	−0.043***	−0.115***	−0.042***	−0.025***	−0.029**	−0.009	−0.103***	−0.018*
migration	(0.003)	(0.010)	(0.011)	(0.019)	(0.010)	(0.005)	(0.010)	(0.017)	(0.018)	(0.007)
Adj. R^2	0.180	0.289	0.307	0.309	0.161	0.175	0.195	0.218	0.275	0.193
Obs.	800,499	617,538	552,234	780,129	341,747	787,531	652,180	523,251	621,221	228,270

Standard errors are clustered at the state level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *state* fixed effects.
†: Only non-Jewish immigrants born in the specific country of origin are included in the sample (Jewish dummy is omitted in the model).
‡: Only non-Jewish immigrants born in the specific country of origin are included in the dependent variable (Jews are not considered in enclaves).

I conclude the discussion of main results in Table 16, in which I use the percentage of immigrants born in the same country of origin—rather than the number of immigrants—to measure the size of country-of-birth enclaves. Results show qualitatively similar patterns of within-group segregation. While not reported here, I observe similar results when using immigrants speaking the same mother tongue to measure enclaves (i.e., language enclaves). These results suggest the main findings of this paper are robust to changes to specifications.

5.3 Measurement Errors

One potential statistical issue of the empirical strategy is that it might be built on misclassification of surname-based ethnicity, which could make the estimates biased.

In Table 17 I redo the empirical analysis based on the sub-sample that only contains surnames that are classified by the deterministic algorithm, but not the Bayes classifier. In other words, I only keep surnames that are less likely to be associated with measurement errors. I also exclude Jews in both the sample and the dependent variable (i.e., country-of-birth enclaves), as measurement errors mainly occur in the Bayes algorithm that classifies

Table 17: Measurement Errors

	The Number of Immigrants Born in the Same Country of Origin, ED					
	(1)	(2)	(3)	(4)	(5)	(6)
	1920 Sample, (1) - (3)			1930 Sample, (4) - (6)		
Sample:	Germany†‡	Poland†‡	Russia†‡	Germany†‡	Poland†‡	Russia†‡
Average:	75.082	410.024	163.856	75.237	214.328	87.174
German-		−145.285***	16.757**		−101.258***	29.536**
name		(18.972)	(6.137)		(9.995)	(9.971)
Polish-	2.329		−26.620	−8.127*		−14.567*
name	(5.015)		(25.590)	(3.134)		(6.806)
Russian-	−11.676***	−107.876***		−11.197***	−78.446***	
name	(1.144)	(11.920)		(1.736)	(13.317)	
Yrs. since	−0.184***	−0.399	−1.184***	−0.397***	−0.132	−0.686**
migration	(0.043)	(0.298)	(0.301)	(0.080)	(0.336)	(0.217)
Adj. R^2	0.330	0.700	0.270	0.349	0.387	0.299
Obs.	312,855	127,592	79,308	312,994	136.992	56,813

Standard errors are clustered at the state level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *state* fixed effects.
†‡: Jews are excluded in both the dependent variable (country-of-birth enclaves) and the sample.

German, Polish, and Russian surnames. Table 17 shows that the results based on this subsample are very similar to the main results reported earlier, suggesting that measurement errors should not affect the empirical conclusion of this paper.

5.4 Additional Tests

The main results of this paper present empirical evidence of within-group ethnic segregation. However, the results are not necessarily causal: if immigrant enclave residence was actually driven by some geographic characteristics, then ethnic segregation might simply be the "coincidence", but not specifically related to ethnicity.

To study this, in Table 18 I include county fixed effects in the regression. In the history of U.S. immigration, some counties have particularly been "gateway counties" (e.g., New York County for Jews), and thus *state* might be too large to serve as the geographic control. Including county fixed effects could account for the potential effects of such "gateway counties". Table 18 presents the results based on county fixed effects. Compared with main results reported earlier, the magnitudes of most coefficients do become smaller, suggesting that county-level geographic characteristics might indeed affect immigrants' settlements. However, the conclusion of within-group ethnic segregation remains true.

Table 18: Within-Group Ethnic Segregation, County Fixed Effects

	The Number of Immigrants Born in the Same Country of Origin (The Size of the Country-of-Birth Enclave), ED								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample:	Germany	Poland	Poland†	Poland‡	Poland†‡	Russia	Russia†	Russia‡	Russia†‡
A. 1920:									
German-		−56.229***	−52.286***	−50.448***	−53.208***	−10.131	3.901	39.953***	8.844**
name		(7.984)	(6.637)	(6.784)	(6.745)	(5.189)	(2.370)	(10.096)	(3.311)
Polish-	−0.806					−42.820**	−13.808	17.524*	−0.738
name	(2.306)					(9.454)	(7.597)	(8.202)	(8.010)
Russian-	−2.849***	−40.272***	−37.930***	−35.232***	−39.196***				
name	(0.482)	(6.497)	(5.328)	(5.576)	(5.383)				
Jewish	−9.325***	−93.502		−175.324***		135.219***		−93.363***	
	(1.885)	(48.339)		(31.546)		(18.400)		(11.612)	
Adj. R^2	0.468	0.705	0.724	0.712	0.725	0.521	0.427	0.326	0.369
Obs.	800,499	617,538	552,234	617,538	552,234	780,129	341,747	780,129	341,747
B. 1930:									
German-		−46.230***	−39.183***	−41.163***	−41.846***	−3.772	7.743	21.797***	8.268*
name		(5.209)	(4.599)	(5.188)	(5.037)	(3.097)	(4.050)	(5.461)	(4.211)
Polish-	−7.040**					−27.644**	−16.531**	−3.276	−10.224**
name	(2.325)					(8.671)	(5.577)	(2.553)	(3.475)
Russian-	−3.941***	−30.578***	−25.838***	−26.373***	−28.196***				
name	(0.806)	(3.398)	(3.101)	(3.206)	(3.217)				
Jewish	−9.162***	−57.407		−150.451***		92.446***		−41.139***	
	(2.314)	(30.745)		(27.132)		(13.998)		(5.929)	
Adj. R^2	0.462	0.427	0.471	0.469	0.469	0.533	0.469	0.316	0.396
Obs.	787,531	652,810	523,251	652,810	523,251	621,221	228,270	621,221	228,270

Standard errors are clustered at the county level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *county* fixed effects.
†: Only non-Jewish immigrants born in the specific country of origin are included in the sample (Jewish dummy is omitted in the model).
‡: Only non-Jewish immigrants born in the specific country of origin are included in the dependent variable (Jews are not considered in enclaves).

I further discuss two specific types of geographic characteristics that might drive ethnic segregation. In Table 19, I examine the heterogeneous effects by urban status. I first focus on the urban sample in Panel A. The empirical conclusion based on the urban sample is very similar to that reported in Section 5.2. In Panel B I repeat the exercise in the rural sample. The results are slightly different, in the sense that coefficients of the Jewish dummy shown in Column 1 and 4 for the 1920 sample (and Column 6 and 9 for the 1930 sample) are insignificant. However, note that only 3% of Jews in the sample lived outside urban areas, and thus Table 19's results might be due to small sample bias. Except for this, the results based on the rural sample are generally consistent with the main results.

I then turn to discuss local labor market characteristics at the ED level. Some ethnic groups were highly concentrated in not only small geographic areas, but also small occupation categories. Hence, ethnic segregation might actually be the consequence of occupational agglomeration, rather than immigrants' settlement patterns. To study this, in

Table 19: Heterogeneous Effects by Urban Status

	The Number of Immigrants Born in the Same Country of Origin (The Size of the Country-of-Birth Enclave), ED									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	1920 Sample, (1) - (5)					1930 Sample, (6) - (10)				
Sample:	Germany	Poland	Poland†‡	Russia	Russia†‡	Germany	Poland	Poland†‡	Russia	Russia†‡
A. Urban:										
Average:	81.601	424.066	449.593	347.753	163.040	85.107	226.821	228.615	259.915	86.818
German-		−85.792***	−90.011***	−1.790	17.619***		−63.755***	−68.124***	3.386	17.057**
name		(15.603)	(15.691)	(8.603)	(4.242)		(7.776)	(9.702)	(7.359)	(5.688)
Polish-	−0.748			−71.950**	−11.911	−9.731***			−51.122***	−15.654***
name	(3.974)			(21.622)	(17.043)	(2.389)			(10.686)	(3.999)
Russian-	−5.447***	−68.614***	−73.066***			−6.273***	−44.482***	−47.999***		
name	(0.900)	(11.679)	(11.810)			(0.935)	(5.538)	(6.292)		
Jewish	−13.404**	−123.526***		178.704***		−13.172**	−73.594*		133.314***	
	(3.638)	(24.636)		(15.467)		(3.731)	(30.770)		(16.164)	
Adj. R^2	0.292	0.659	0.675	0.433	0.262	0.299	0.331	0.355	0.412	0.261
Obs.	531,100	509,878	447,257	686.483	259.587	563,149	558,366	431,813	557,713	172,045
B. Rural:										
Average:	55.157	130.559	131.319	101.578	100.038	38.762	90.152	90.287	47.189	43.585
German-		−29.493***	−31.130***	14.223*	16.267**		−14.358***	−14.814***	7.597**	6.491***
name		(6.347)	(6.493)	(6.845)	(5.799)		(2.054)	(2.192)	(2.324)	(1.491)
Polish-	0.451			−3.796	1.432	1.637			−3.204	−2.513
name	(1.490)			(3.982)	(1.974)	(0.978)			(2.549)	(2.506)
Russian-	−3.154***	−26.581***	−28.470***			−2.801***	−11.550***	−12.034***		
name	(0.736)	(6.804)	(7.175)			(0.563)	(1.785)	(1.925)		
Jewish	0.326	−39.454***		26.647		0.982	−16.346***		13.162	
	(0.888)	(8.129)		(20.887)		(1.071)	(2.955)		(7.652)	
Adj. R^2	0.384	0.386	0.388	0.342	0.367	0.443	0.424	0.427	0.370	0.363
Obs.	269,399	107,660	104,977	93,646	82,160	224,382	94,444	91,438	63,508	55,865

Standard errors are clustered at the state level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *state* fixed effects.
†: Only non-Jewish immigrants born in the specific country of origin are included in the sample (Jewish dummy is omitted in the model).
‡: Only non-Jewish immigrants born in the specific country of origin are included in the dependent variable (Jews are not considered in enclaves).

Table 20 I reestimate Equation 4, but now with the inclusion of several ED-level labor market variables, such as average occupational scores by group, and average age. Compared with the main results of this paper, Table 20 shows that the magnitudes of the coefficients are smaller in almost all columns, suggesting that local labor market characteristics at the ED level might indeed partially account for within-group ethnic segregation. However, the qualitative pattern shown in Table 20 is still consistent with the main results reported earlier: within-group ethnic segregation was significant in the early twentieth century U.S. even after taking possible effects of occupational agglomeration into consideration.

6 Conclusion

Many studies consider *ethnicity* or *ethnic origin* as the equivalent term for the country of birth (e.g., Fairlie and Meyer, 1996). This is a safe assumption if a sending country sends demographically homogeneous ethnic group. In the age of mass migration, however, most immigrants were from Europe, and many European countries had high degrees of

Table 20: Segregation and Local Labor Market (ED) Characteristics

	The Number of Immigrants Born in the Same Country of Origin (The Size of the Country-of-Birth Enclave), ED								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample:	Germany	Poland	Poland†	Poland‡	Poland†‡	Russia	Russia†	Russia‡	Russia†‡
A. 1920:									
German-		−54.247***	−54.519***	−51.413***	−56.603***	9.154	19.714***	48.791**	20.501***
name		(11.141)	(10.591)	(10.168)	(11.453)	(6.542)	(2.904)	(15.261)	(3.553)
Polish-	0.299					−71.522***	−36.039*	8.182	−16.809
name	(2.791)					(17.269)	(14.983)	(9.826)	(13.007)
Russian-	−4.623***	−57.249***	−58.238***	−53.584***	−60.220***				
name	(0.787)	(8.021)	(7.998)	(7.509)	(8.198)				
Jewish	−9.824**	−92.962***		−195.089***		149.944***		−89.578***	
	(3.422)	(10.521)		(27.470)		(15.801)		(16.661)	
Adj. R^2	0.335	0.681	0.699	0.685	0.697	0.524	0.386	0.287	0.325
Obs.	800,499	617,538	552,234	617,538	552,234	780,129	341,747	780,129	341,747
B. 1930:									
German-		−41.827***	−38.934***	−39.918***	−43.851***	−7.513	18.568***	28.213***	15.850***
name		(5.249)	(3.683)	(4.744)	(5.564)	(5.916)	(3.697)	(3.948)	(4.208)
Polish-	−5.781**					−45.115***	−25.586***	−6.701*	−14.818**
name	(1.705)					(8.468)	(4.892)	(3.129)	(3.016)
Russian-	−5.440***	−34.300***	−32.443***	−31.517***	−35.791***				
name	(0.943)	(3.494)	(3.725)	(3.323)	(3.636)				
Jewish	−10.643**	−39.836		−146.678***		119.372***		−41.222***	
	(3.762)	(21.535)		(18.016)		(18.144)		(5.184)	
Adj. R^2	0.359	0.441	0.481	0.476	0.475	0.484	0.354	0.209	0.273
Obs.	787,531	652,810	523,251	652,810	523,251	621,221	228,270	621,221	228,270

Standard errors are clustered at the state level and are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.
In all regressions, I control for all individual characteristics introduced in Section 4.1 and *state* fixed effects.
†: Only non-Jewish immigrants born in the specific country of origin are included in the sample (Jewish dummy is omitted in the model).
‡: Only non-Jewish immigrants born in the specific country of origin are included in the dependent variable (Jews are not considered in enclaves).

ethnic and cultural diversity, leading to possible ethnic differences within one immigrant group defined based on the country of birth. In this paper, I examine whether such ethnic differences could lead to within-group ethnic segregation at the county level in the U.S.

Focusing on first-generation immigrants who reported Germany, Poland, or Russia as the birthplace in the 1920 and 1930 census, I start the empirical analysis by creating an ethnicity variable based on the linguistic origin of the surname. This is based on findings of human biology and ethnography that the surname origin is highly associated with ethnicity (e.g., Guglielmino et al., 2000; Schramm et al., 2012). I use both the deterministic and probabilistic algorithm to classify surname-based ethnicity for all male immigrants born in Germany, Poland, or Russia in the full-count 1920 and 1930 U.S. census.

I then use this ethnicity variable to calculate the dissimilarity index of ethnic segregation. Results suggest that the degree of within-group ethnic segregation was high; in some cases, ethnic segregation within the immigrant group defined based on country of birth could be even higher than immigrant-native segregation. Within-group ethnic segregation could also be higher than segregation by the country of birth within one surname-based ethnic population. Subsequently, I study ethnic differences in "country-of-birth enclave residence" within a regression framework. Results show significant evidence of ethnic segregation at the enumeration district (ED) level, and ethnic minorities generally lived in EDs with much fewer compatriots. The results are robust to changes to samples and specifications. While the country of birth is still a key variable in immigration studies, the empirical conclusion of this paper highlights the importance of considering heterogeneity in ethnicity in an immigrant group defined based on the country of birth.

References

[1] Abramitzky, Ran, and Leah Platt Boustan. 2017. "Immigration in American Economic History." *Journal of Economic Literature*, 55(4), 1311 - 1345.

[2] 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.

[3] Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2016. "Cultural Assimilation during the Age of Mass Migration." NBER Working Paper No. 22381.

[4] Altonji, Joseph G., and David Card. 1994. "The Effects of Immigration on the Labor Market Outcomes of Less-skilled Natives." in *Immigration, Trade, and the Labor Market*, edited by John M. Abowd and Richard B. Freeman. Chicago: University of Chicago Press.

[5] Ambekar, Arunug, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. 2009. "Name-Ethnicity Classification from Open Sources." *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 49 - 58.

[6] Arai, Mahmood, and Peter Skogman Thoursie. 2009. "Renouncing Personal Names: An Empirical Examination of Surname Change and Earnings." *Journal of Labor Economics*, 27(1), 127 - 147.

[7] Bahlow, Hans. 2002. *Dictionary of German Names*. Madison: Max Kade Institute.

[8] Bartel, Ann P. 1989. "Where Do the New U.S. Immigrants Live?" *Journal of Labor Economics*, 7(4), 371 - 391.

[9] Bleakley, Hoyt, and Aimie Chin. 2004. "Language Skills and Earnings: Evidence from Childhood Immigrants." *Review of Economics and Statistics*, 86(2), 481 - 496.

[10] Biavaschi, Constanza, Corrado Giuliett, and Zahra Siddique. 2017. "The Economic Payoff of Name Americanization." *Journal of Labor Economics*, 35(4), 1089 - 1116.

[11] Boustan, Leah Platt. 2007. "Were Jews Political Refugees or Economic Migrants? Assessing the Persecution Theory of Jewish Emigration, 1881-1914." in Timothy J. Hatton, Kevin H. O'Rourke, and Alan M. Taylor, eds., *The New Comparative Economic History: Essays in Honor of Jeffrey G. Williamson*. cambridge: The MIT Press.

[12] Corrisin, Stephen D. 1990. "Language Use in Cultural and Political Change in Pre-1914 Warsaw: Poles, Jews, and Russification." *Slavonic and East European Review*, 68(1), 69 - 90.

[13] Chibnik, Michael. 1991. "Quasi-Ethnic Groups in Amazonia." *Ethnology*, 30(2), 167 - 182.

[14] Crowley, Terry, and Claire Bowern. 2010. *An Introduction to Historical Linguistics*. New York: Oxford University Press.

[15] Cutler, David M, Edward L. Glaeser, and Jacob L. Vigdor. 2008. "Is the Melting Pot Still Hot? Explaining the Resurgence of Immigrant Segregation." *Review of Economics and Statistics*, 90(3), 478 - 497.

[16] Dubin, Lois C. 1999. *The Port Jews of Habsburg Trieste: Absolutist Politics and Enlightenment Culture*. Stanford: Stanford University Press.

[17] Duncan, Otis Dudley, and Stanley Lieberson. 1959. "Ethnic Segregation and Assimilation." *American Journal of Sociology*, 64(4), 364 - 374.

[18] Edin, Per-Anders, Peter Fredriksson and Olof Åslund. 2003. "Ethnic Enclaves and the Economic Success of Immigrants: Evidence from a Natural Experiment." *Quarterly Journal of Economics*, 118(1), 329 - 357.

[19] Fairlie, Robert W., and Bruce D. Meyer. 1996. "Ethnic and Racial Self-Employment Differences and Possible Explanations." *Journal of Human Resources*, 31(4), 757 - 793.

[20] Foley, C. Fritz, and William R. Kerr. 2013. "Ethnic Innovation and U.S. Multinational Firm Activity." *Management Science*, 59(7), 1529 - 1544.

[21] Fucilla, Joseph Guerin. 1998. *Our Italian Surnames*. Baltimore: Genealogical Publishing Company.

[22] Fussell, Elizabeth. 2014. "Warmth of the Welcome: Attitudes Toward Immigrants and Immigration Policy in the United States." *Annual Review of Sociology*, 40, 479 - 498.

[23] Gordon, Milton M. 1964. *Assimilation in American Life: The Role of Race, Religion, and National Origins*. New York: Oxford University Press.

[24] Greenwood, Michael J., and Zachary Ward. 2015. "Immigration Quotas, World War I, and Emigrant Flows from the United States in the Early 20th Century." *Explorations in Economic History*, 55, 76 - 96.

[25] Guglielmino, C.R., G., Zei, and L.L. Cavalli-Sforza. 2000. "Genetic and Cultural Transmission in Sicily as Revealed by Names and Surnames." *Human Biology*, 63(5), 607 - 627.

[26] Haines, Michael R. 2000. "The Population of the United States, 1790 - 1920." In *The Cambridge Economic History of the United States*, eds., Stanley L. Engerman and Robert E. Gallman. New York: Cambridge University Press.

[27] Hoffman, William F. 2001. *Polish Surnames: Origins and Meanings*. Chicago: Polish Genealogical Society of America.

[28] Kriendler, Isabelle T. 1985. *Sociolinguistic Perspectives on Soviet National Languages: Their Past, Present and Future*. Berlin: De Gruyter Mouton.

[29] Mateos, Pablo. 2007. "A Review of Name-Based Ethnicity Classification Methods and their Potential in Population Studies." *Population, Space and Place*, 13(4), 243 - 263.

[30] Monasterio, Leonardo. 2017. "Surnames and Ancestry in Brazil." *PLoS ONE*, 12(5), e0176890. Available online: https://doi.org/10.1371/journal.pone.0176890.

[31] Munshi, Kaivan. 2003. "Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market." *Quarterly Journal of Economics*, 118(2), 549 - 599.

[32] Powers, David M. w. 2011. "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation." *Journal of Machine Learning Technologies*, 2(1), 37 - 63.

[33] Razum, Oliver, Hajo Zeeb, and Seval Akgün. 2001. “How Useful is a Name-Based Algorithm in Health Research among Turkish Migrants in Germany?” *Tropical Medicine and International Health*, 6(8), 654 - 661.

[34] Reaney, Percy H. 2005. *A Dictionary of English Surnames*. New York: Oxford University Press.

[35] Rish, I. 2005. “An Empirical Study of the Naive Bayes Classifier.” *IJCAI workshop on Empirical Methods in AI*. http://ai2-s2-pdfs.s3.amazonaws.com/2825/733f97124013e8841b3f8a0f5bd4ee4af88a.pdf.

[36] Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2017. *Integrated Public Use Microdata Series: Version 7.0* [Machine-readable database]. Minneapolis: University of Minnesota.

[37] Sammartino, Annemarie H. 2010. *The Impossible Border: Germany and the East, 19141922*. Ithaca: Cornell University Press.

[38] Sassler, Sharon. 2005. “Gender and Ethnic Differences in Marital Assimilation in the Early Twentieth Century.” *International Migration Review*, 39(3), 608 - 634.

[39] Schramm, Katherine, David Skinner, and Richard Rottenburg. 2012. *Identity Politics and the New Genetics: Re/Creating Categories of Difference and Belonging*. New York: Berghahn Books.

[40] Steckel, Richard H. 1983. “The Economic Foundations of East-West Migration during the 19th Century.” *Explorations in Economic History*, 20(1), 14 - 36.

[41] Stern, Malcolm H., and Dan Rottenberg. 1998. *Finding Our Fathers: A Guidebook to Jewish Genealogy*. Baltimore: Genealogical Publishing Company.

[42] Sue, Christina A., and Edward E. Telles. 2007. “Assimilation and Gender in Naming.” *American Journal of Sociology*, 112(5), 1383 - 1415.

[43] Unbegaun, Boris Ottokar. 1972. *Russian Surnames*. Oxford: Oxford University Press.

[44] van Nuys, Frank. 2002. *Americanizing the West: Race, Immigrants, and Citizenship, 1890-1930*, Lawrence: University Press of Kansas.

[45] Ward, Zachary. 2017. “Birds of Passage: Return Migration, Self-Selection and Immigration Quotas.” *Explorations in Economic History*, 64, 37 - 52.

[46] Waters, Mary C. 1989. “The Everyday Use of Surname to Determine Ethnic Ancestry.” *Qualitative Sociology*, 12(3), 303 - 324.

[47] Winship, Christopher. 1977. “A Revaluation of Indexes of Residential Segregation.” *Social Forces*, 55(4), 1058 - 1066.

[48] Xu, Dafeng. 2017. “Acculturational Homophily.” *Economics of Education Review*, 59, 29 - 42.