Identifying Ethnic Occupational Segregation: A Machine Learning Approach

Dafeng Xu   Yuxin Zhang*

*Dafeng Xu is at the University of Washington, Evans School of Public Policy & Governance, 4105 George Washington Lane Northeast, Seattle, WA 98105, USA. Email: dafengxu@uw.edu. Yuxin Zhang is at Wayne State University, Mike Ilitch School of Business, 2771 Woodward Avenue, Detroit, MI 48201, USA, and the University of Texas at Austin, McCombs School of Business, 2110 Speedway, Austin, TX 78705, USA. Email: yuxin.zhang@utexas.edu.

Abstract

Many studies consider an immigrant population (defined by birth country) as a whole. This might neglect ethnic heterogeneity within an immigrant population and thus underestimate occupational segregation. We focus on Russian immigrants in the early 20th century U.S.—then a major immigrant population with a high degree of ethnic diversity (consisting of Russian, Jewish, German, and Polish ethnics)—and study occupational segregation by ethnicity. We develop a machine learning ethnicity classification approach in 1930 U.S. census data based on name and mother tongue. Using the constructed ethnicity variable, we show high degrees of occupational segregation by ethnicity within the Russian-born immigrant population in the U.S. For example, Jews, German ethnics, and Polish ethnics were concentrated in trade, agriculture, and manufacturing, respectively. We also find evidence that Russian-born immigrants' labor market outcomes were associated with networks measured by the spatial concentration of co-ethnics—more established ones in particular—but not other ethnic groups.

Keywords: immigration, occupational segregation, ethnicity, network, machine learning

JEL Classification: J1, N3, R2

1 Introduction

Scholars have long observed that U.S. immigrants have unique labor market patterns (e.g., Borjas, 1986; Yuengert, 1995), and different immigrant populations also have their own labor market patterns (e.g., Fairlie and Meyer, 1996). Many studies consider an immigrant population (defined by country of birth) as a whole and use *country of birth* interchangeably with *ethnicity*. However, ethnicity actually means a category of people classified on the basis of a common genealogy/ancestry (Chibnik, 1991) or culture (Constant and Zimmermann, 2008; Constant et al., 2009). Indeed, economists present empirical evidence that some ethnic sub-populations could have unique economic outcomes: for example, Kalnins and Chung (2006) observe that social capital plays a unique role in the U.S. lodging industry among Gujarati immigrants; Moser et al., (2014) find positive effects of the influx of German Jews on R&D in the U.S.; O'Keefe and Quincy (2018) show that Jewish immigration from Russia affected native-born farmers in New Jersey in the early 20th century. Treating birth country and ethnicity as two separate concepts helps us understand important labor economic questions such as occupational segregation by ethnicity, which is driven by ethnic differences in pre-migration human capital accumulation, selective migration, and post-migration ethnic enclave residence.

In this paper, we first present a machine-learning-based algorithm of ethnic classification and then analyze occupational segregation by ethnicity among Russian-born immigrants in the U.S. using 1930 U.S. full-count census data (Ruggles et al., 2019). During the age of mass migration (the late 19th and early 20th century U.S.), the majority of immigrants came from Europe, and some sending countries had high degrees of ethnic diversity,[1] including the Russian Empire. Specifically, Russian-born immigrants were one of the largest immigrant populations in the early 20th century U.S. (Haines, 2000) and had a high degree of ethnic diversity, with a mixture of Russian, Jewish, German, and Polish

[1] Eberhardt (2002) summarizes ethnic majorities and minorities in Eastern Europe in the early 20th century. Haines (2000) presents some examples of ethnic heterogeneity among European immigrants, such as Eastern European Jews who were minorities in their home countries and were more likely to move to the U.S.

ethnic sub-groups (Lieven, 2006).[2]

This paper makes methodological contributions to population economics by developing a machine-learning-based probabilistic method of ethnicity classification using name and language variables in census data. Ethnicity is a useful variable in the economic analysis of immigration and the labor market (e.g., Constant and Zimmermann, 2009), which will be discussed in more detail later. Prior studies use deterministic linkages between names in the sample and name dictionaries to classify ethnicity (e.g., Foley and Kerr, 2013). However, due to low data quality, deterministic classification can only identify ethnicity of less than half of cases in digitized historical census data (Xu, 2019). This paper extends existing strategies by introducing machine learning tools for ethnicity classification.

The substantive topic of this paper adds to the literature of labor economics along multiple dimensions. First, this paper sheds light on the understandings of the origin of within-population occupational segregation. Although the immigrants studied in this paper shared the same country of birth (Russia), there could be huge heterogeneity in occupational patterns by ethnicity among these co-birth immigrants. Such heterogeneity could be originally formulated roughly by three types of reasons. First, there were ethnic differences in human capital characteristics in Russia, when different ethnic groups had unique patterns of educational attainment (Nathans, 2006; Snyder, 2006) and were exposed to school and residential segregation (Nathans, 2002). These differences further led to *pre-migration* occupational segregation by ethnicity in Russia due to differences in the requirements of human capital and geographic distributions of occupations. Second, in early 20th century Russia, various economic and political factors related to ethnicity—including ethnic conflicts, hate crimes against minorities, anti-minority populism, education quota laws, restrictions on work authorization, and ethnic differences in residential locations[3]—led to selection on emigration

[2] In this paper, we use the term *within-population* to describe "within the Russian-born immigrant population in the U.S." We use the term *Russian-born* immigrants to describe U.S. immigrants born in Russia, regardless of ethnicity. We use the term *Russian ethnic* (or Jewish, German, Polish ethnic) immigrants to describe Russian-born immigrants of Russian (or Jewish, German, Polish, respectively) ethnicity.

[3] See summaries of ethnicity-related laws, policies, and conflicts in Eastern Europe by Petersen (2002), and specifically in Russia by Klier (1995), Dowler (2000), Nathans (2002), and Thaden (2014).

by ethnicity. Third, and relatedly, ethnic populations were exposed to Russia's political and economic reforms at different levels, including heterogeneous effects of the Emancipation Reform (Wetherell and Plakans, 1999) and Stolypin Reform (Chernina et al., 2014) by ethnicity, which further led to ethnic differences in migration to the U.S.

Furthermore, measuring immigrants' ethnicity sheds light on the analysis of immigrants in the labor market. First, in the 20th century, Russian-born immigrants contributed to the U.S. economy in multiple industries, such as trade (Simon, 1997; Polland and Soyer, 2013), manufacturing (Zunz, 2000; Pacyga, 2003), and agriculture (O'Keefe and Quincy, 2018). The multiple dimensions of Russian-born immigrants' contributions could be related to ethnic occupational segregation, which could further trace back to pre-migration ethnic differences in educational attainment and employment. This presents linkages between pre-migration characteristics and post-migration outcomes, which is an important topic in immigration studies (e.g., Hirsch et al., 2014; Blau and Kahn, 2015; Polavieja, 2015). Second, Abramitzky et al. (2014) find that economic conditions of the home country were closely associated with post-migration economic status among immigrants, but immigrants born in Russia—then a relatively poor country—had almost the highest occupation-based earnings, which could be potentially explained by occupational segregation and selective migration by ethnicity. Third, researchers observe that U.S. immigrants benefit from ethnic social networks.[4] The economic effects of networks might also exist in this paper's context, and the first step to study this is to design an accurate measure of ethnicity.

We classify ethnicity based on two types of ethnicity widely used in anthropology and history, i.e., language-based ethnicity (e.g., Duranti, 1997) and name-based ethnicity (e.g., Waters, 1989; Chibnik, 1991). We construct language-based ethnicity based on mother tongues surveyed in the census. We construct name-based ethnicity by employing a naïve

[4] While prior studies find conflicting evidence of effects of ethnic enclave residence, the consensus is that ethnic enclaves have positive impacts after considering selection (Borjas, 1994; Edin et al., 2003; Cutler et al., 2008; Senik and Verdier, 2011), which was common in the early 20th century (Xu, 2019). Using survey data that provide information on ethnicity, Munshi (2003) and Kalnins and Chung (2006) find the economic effects of social networks formed within ethnic *sub-groups* in a co-birth immigrant population. We follow these studies in a new data setting in which ethnicity information are not provided by census respondents.

Bayes classifier (Rish, 2005) to measure linguistic origins of names based on ethnic name dictionaries (Xu, 2019). The "overall" ethnicity is then calculated based on the combination of two types of ethnicity. Results in different test data show the robustness of our method. Although deterministic ethnicity classification has been widely studied in social sciences (e.g., Mateos, 2007), including in economics (e.g., Foley and Kerr, 2013; Zhang, 2016), our method allows fuzzy classification in low-quality data (due to transcription errors, typos, and name changes), such as early census data used in this paper.

Consistent with historical findings (e.g., Simon, 1997; Nathans, 2002), we show that Jews were the majority Russian-born ethnic group in the U.S. Applying our ethnicity measure, we observe significant occupational segregation by ethnicity among Russian-born immigrants. For example, Jews were more likely to work in trade, German ethnics were more likely to work in agriculture, and Polish ethnics were more likely to work in manufacturing. The results remain unchanged when controlling for individual characteristics. We further present case studies and argue that ethnic occupational segregation among Russian-born immigrants in the U.S. observed in the empirical analysis appears to be similar to that in Russian cities in the late 19th and early 20th century.

We then explore the relationship between the spatial concentration of co-ethnics and occupational outcomes. This is motivated by findings that immigrants' outcomes are influenced by networks measured by the concentration of co-birth immigrants (e.g., Edin et al., 2003; Patacchini and Zenou, 2012; Kalfa and Piracha, 2018). Baseline OLS estimates show that the concentration of co-ethnics—in particular, the more established ones (consistent with findings of Munshi, 2003)—was positively correlated with employment status, earnings, and occupational standings; the concentration of other ethnic groups had weaker or null effects. These baseline results are consistent with instrumental variable (IV) estimations, in which we use historical settlements of the particular ethnic groups as IVs.

The remainder of this paper is organized as follows. Section 2 introduces the historical background. Section 3 describes data and methods of ethnicity classification. Focusing on

U.S. census data, Section 4 studies occupational segregation by ethnicity among Russian-born immigrants and the spatial mechanisms. Section 5 concludes.

2 Historical Background

2.1 Ethnic Sub-Groups among Russian-Born Immigrants

In anthropology, ethnicity is defined as a category of people who have the same ancestry or culture (e.g., Chibnik, 1991; Guglielmino et al., 2000; Peoples and Bailey, 2010). In other words, individuals within an ethnic group share similar biological and cultural characteristics. Such similarities in biological (e.g., genes, see Mountain and Risch, 2004) and sociocultural (e.g., education, see Carlton and Weiss, 2001) patterns generate ethnic-specific outcomes. For example, Spielman et al. (2007) find that different ethnic populations have genetic variants that trace back to geographic divisions of ethnic ancestries that have long-run effects on genetic makeup and phenotype, which lead to variations in genetic expressions and further cause ethnic differences in prevalence of genetic diseases (e.g., diabetes and cardiovascular diseases). Botticini and Eckstein (2007) find Jews' self-selection into urban and skilled occupations in the first millennium even in regions with no restrictions on their economic activities. More broadly, Connor (1993) find that ethnic identities lead to the belief of *ethnonationalism* that formulates the political foundation of discrimination and oppression against minorities, including in the U.S. (Ngai, 1999).

Most surveys do not ask questions about ethnicity. However, it is possible to construct proxies for ethnicity using various demographic variables. One strand of literature focuses on names and uses linguistic origins of names to classify ethnicity (Mateos, 2007). This is based on the findings that names reflect biological and cultural transmission. In ancient times, ethnic ancestries were divided in space (Mateos, 2014) and developed idiosyncratic linguistic patterns (Crowley and Bovern, 2010), based on which many ethnic surnames evolved (Chibnik, 1991), and thus names reflect ethnic-specific genetic features

(Guglielmino et al., 2000) through genetic variants by ethnicity (Mountain and Risch, 2004; Spielman et al., 2007). Moreover, as names are related to early language development that follows ethnic-specific paths, many ethnic populations consider names as cultural identities (Mateos, 2014) that can be transmitted across generations (Waters, 1989; Monasterio, 2017). Based on the above rationale, economists classify ethnicity by linking names in their data with ethnic name dictionaries. For example, Foley and Kerr (2013) and Zhang (2016) study ethnic differences in patent records and fair lending risks based on name-based ethnicity classification. Relatedly, economists observe name-based discrimination (e.g., Bertrand and Mullainathan, 2004; Drydakis, 2011; Oreopolous, 2011; Rubinstein and Brenner, 2014) due to the relationship between names and ethnicity. Compared with these studies, our approach allows fuzzy classification, which is needed in the context of this paper where historical census records have relatively low data quality. We will discuss fuzzy classification in detail in Section 3.1.

Another two useful variables are mother tongue and religion. In anthropology, language is used to identify race and ethnicity (Duranti, 1997) and is widely usable in demographic research because most social surveys ask questions about the mother tongue. Religion is another proxy for ethnicity because many religions are ethnic-specific within a country (Eriksen, 2002); however, many surveys (especially surveys in the U.S.) do not ask questions about religion. Note that in practice, the classification algorithm is more effective if ethnicity is jointly determined by the above variables. The main reason is that many ethnic minorities might socially and culturally assimilate into the mainstream society before migrating to the U.S. For example, ethnic minorities were required to learn Russian in school since the late 19th century in Imperial Russia (Dowler, 2000; Sammartino, 2010), which could make language-based ethnicity less reliable; on the other hand, name Russification was less common than language Russification (Thaden, 2014). In general, having multiple measures of ethnicity could reduce measurement errors.

Many prior studies simply consider ethnicity to be an equivalent term as country of birth (e.g., Fairlie and Meyer, 1996; Bleakley and Chin, 2004). This is not a valid assumption if there are multiple significant ethnic groups within a co-birth immigrant population, such as Russian-born immigrants during the age of mass migration: in the early 20th century, there were over one million of immigrants moving to the U.S. from Russia, which mainly consisted of Russian, Jewish, German, and Polish ethnics.

[Insert Table 1 here.]

The 1897 Russian Empire Census (Central Statistical Committee of Russia, 1897) recorded 125,640,021 people in Imperial Russia. While the 1897 Russian census did not survey ethnicity, it asked questions about the mother tongue and religious affiliation. Panel A of Table 1 shows census results by language and religion. Under the language classification, 44.31% (Russian language only) or 66.80% (including Ukrainian and Belorussian language, i.e., pan-Russian language) of the full population were Russian/pan-Russian language speakers; 4.03%, 1.43%, and 6.31% of the population spoke Yiddish (Jewish language in Eastern Europe), German, and Polish, respectively. Under the religion classification, 69.34% of the full population were affiliated with Eastern Orthodox; 4.15%, 2.84%, and 9.13% of the population were affiliated with Jewish, Lutheran, and Roman Catholics, respectively. Note that some of the large minority groups (e.g., Latvians and Lithuanians) also contributed to the religious populations. In general, Panel A of Table 1 shows that in 1897 Russian census data, the ethnic composition classified by language appears to be consistent with that classified by religious affiliation. However, most of the non-Orthodox religious populations were larger than the corresponding language populations, as linguistic assimilation was usually faster and more common than religious assimilation (e.g., Nathans, 2006; Snyder, 2006).

2.2 Migration from Russia to the U.S.

Russia was among the top sending countries of immigrants to the U.S. in the late 19th and early 20th century. Panel B of Table 1 summarizes major sending countries of immigrants.

Immigration from Russia was rare in 1880, especially compared to immigration from "old source countries" such as Germany and Ireland (Haines, 2000). The number of immigrants born in Russia rose significantly in the early 20th century, and in particular, the number of Russian-born immigrants exceeded 1 million soon afterwards. There was a sharp decline in the number of Russian-born immigrants after 1920 due to return migration (Ward, 2017) and immigration restriction laws in 1921 and 1924 (Ngai, 1999), but by 1940, there were still more than 1 million Russian-born immigrants in the U.S.

Panel A of Table 1 shows that the majority of the population in Russia was Russian ethnics, and the largest minority group was the Turkish-Tatar group (by language) or Muslims (by religion). Although the U.S. census did not survey religion, its questionnaire about mother tongue shows that the ethnic composition of the population in Russia was different from that of Russian-born immigrants in the U.S. in the early 20th century. Panel C of Table 1 shows that more than 60% of Russian-born immigrants in the U.S. spoke Yiddish and Hebrew, two Jewish languages. There were also disproportionately more German speakers but disproportionately fewer Pan-Russian language speakers and Polish speakers among Russian-born immigrants. In addition, several large minority groups back in Russia (e.g., Turkish-Tatar) were almost negligible in the U.S. In fact, approximately 98% of all Russian-born immigrants in the U.S. belonged to one of the four major language groups listed in Table 1, and none of the other language groups contributed to more than 0.6% of the Russian-born immigrant population in the U.S.

The inconsistency between the ethnic composition in Russia and Russian-born immigrants in the U.S., as shown in Table 1, suggests possible selection on migration due to several types of reasons. First, selective migration could be driven by ethnicity-related economic and political factors in Russia, including ethnic conflicts, hate crimes against minorities, anti-minority populism, discrimination laws (e.g., education quotas and work restrictions), and residential segregation. Overall, hostility against ethnic minorities was common among ordinary Russian ethnics in the late 19th and early 20th century, and led to mass emigration of ethnic minorities. At that time, anti-Jewish pogroms occurred frequently in Russia (Naimark, 2002; Nathans, 2002, 2006; Boustan, 2007). Poles were also discriminated against (Snyder, 2006). While historically German ethnics enjoyed privileged positions, especially in Baltic states, hostility towards Germans began to rise in Russia in the late 19th and early 20th century due to a decline in Germany-Russia relations (Sammartino, 2010). Hostility also appeared at the governmental level: in the 1880s, a series of discrimination laws on education quotas and work restrictions against Jews (e.g., May Laws in 1882 and *Numerus Clausus* in 1887) were passed and enacted (Nathans, 2002). While German and Polish ethnics were less restricted, the passage of Russification policies led to forced assimilation of these minorities (Dowler, 2000; Thaden, 2014). In addition, ethnic minorities were spatially segregated in Russia. Jews were particularly isolated both within and beyond the *Pale of Settlement*[5] (Klier, 1995; Nathans, 2002), and other ethnic minorities were also moderately segregated. Panel A of Table 2 presents an example of residential segregation of ethnic minorities in St. Petersburg.

[Insert Table 2 here.]

Relatedly, different ethnic populations responded to Russia's political and economic reforms differently, which further led to selection on migration. Specifically, the Emancipation Reform in the 1860s and Stolypin Reform in the early 20th century had great influences on Russia's society and economy, and Russian ethnics were more affected by these major reforms. First, the timing of the emancipation of serfs varied by region: in Poland and Baltic provinces where the majority of Polish and German ethnics resided in, serfdom was abolished in the early 19th century (Wetherell and Plakans, 1999; Snyder, 2006), several decades earlier than that in central Russia. Second, through improvements in property rights, the Stolypin Reform generated migration effects and drove mobility from Europe to Asia (Chernina et al., 2014), and internal migration within Russia mainly originated from

[5] The Pale of Settlement was a geographic region in Western Russia, in which Jewish permanent residency was allowed. Beyond the Pale of Settlement, Jewish residency was mostly forbidden.

central Russia where the majority population was Russian ethnics (Treadgold, 1957).

Finally, an important reason behind self-selection on migration was the capacity of emigration and the relevant knowledge (Haines, 2000), which were related to the population geography of Imperial Russia. In general, as Poland, Ukraine, and the Baltic region (in which the proportion of Polish, German, and Jewish ethnics were higher) were closer to Western Europe, ethnic minorities had better access to emigration and were also more knowledgeable about moving to the rest of the world. Furthermore, ethnic minorities were disproportionately more likely to reside in urban areas (Lieven, 2002), which was also related to the ease of migration.

[Insert Figure 1 here.]

Figure 1 shows geographic distributions of language groups in Russia based on 1897 census data. Two sub-figures in row 1 show the distribution of the Russian-speaking and Polish-speaking population, respectively. Assuming ethnicity could be reflected by mother tongue, Russian ethnics mainly resided in regions that become contemporary Russia, while Polish ethnics mainly resided in today's Poland and western Ukraine. Row 2 focuses on two Germanic languages: Yiddish and German. Most Yiddish speakers (essentially Jews) and German speakers mainly resided in today's Poland, western Ukraine, and the Baltic region. The majority of Jews resided in the *Pale of Settlement* and were generally only allowed to live in cities (Simon, 1997). The majority of German ethnics resided in Baltic region, the traditional settlements of Baltic Germans (Sammartino, 2010). Row 3 focuses on Turkic language groups and specifically Chuvash speakers. Most Turkic speakers resided in central Asia and Far East. These figures show ethnic differences in geographic distributions in Russia, which potentially led to geography-related selection on migration.

The above analysis suggests three points about how geography drove self-selection on migration. First, the geographic distributions of Russian citizens determined self-selection on migration in terms of ethnicity. Specifically, three minority groups of interests—Jews, German ethnics, and Polish ethnics—were disproportionately concentrated in western Russia and were thus more likely to migrate. Second, and relatedly, the above analysis explains why there were disproportionately fewer Russian ethnics and Turkish-Tatar ethnics—two significant groups that constituted nearly 80% of the population in Imperial Russia—among Russian-born immigrants in the U.S. Finally, the above geographic pattern further suggests different degrees of self-selection on migration *within* each ethnic group: from geographic perspectives, the processes of migration from Russia to the U.S. were more selected among Russian ethnics as the proportion of Russian ethnics was relatively lower in western regions of Russia, and Russian ethnics were relatively less likely to live in cities.

In addition to the above points, a particular reason for the large Russian-born Jewish population in the U.S. was the low return migration rate. In general, return migration was common among European immigrants (e.g., Bandiera et al., 2013; Abramitzky et al., 2014). However, almost all Jewish immigrants, including Russian-born Jews, remained in the U.S. after arrival (Ward, 2017). This partially explains the disproportionate presence of Jews in the Russian-born immigrant population in the early 20th century U.S.

2.3 Occupational Patterns among Russian-Born Immigrants

To understand ethnic heterogeneity in occupational patterns among Russian-born immigrants in the U.S., it is useful to first investigate occupational segregation by ethnicity in Imperial Russia. Serfdom was officially abolished after the Emancipation Reform of 1861, but the majority of Russian citizens still lived in rural areas: the 1897 Russian census shows that only 13% of the Russian population resided in cities, and many Russian, Polish, and German ethnics worked in agriculture. The exception was the Jewish population (Simon, 1997): Jews were traditionally not permitted to purchase land in Russia until the early 19th century, and the right to purchase land was again prohibited in the late 19th century following May Laws (Nathans, 2006). As a result, 82% of Jews lived in cities (49%) or small towns (33%), and less than 4% of Jews worked in agriculture. The majority of Jews worked in commerce (39%) and crafts and industry (35%) in Russia (Simon, 1997).

Panels B and C of Table 2 present two specific examples of ethnic occupational patterns in St. Petersburg, then the capital of Russia. Panel B shows that there were a large number of Roman Catholic (related to Polish ethnics) and Jewish lawyers and apprentices lawyers in St. Petersburg in the late 19th century, although St. Petersburg was not a major city of residence for Jews and Polish ethnics in Imperial Russia (see Figure 1). Panel C shows that Jews in St. Petersburg's manufacturing sector were disproportionately more likely to be managers or self-employed and were less likely to be workers.

While there has been no general discussion on ethnic occupational patterns among Russian-born immigrants in the U.S. due to the lack of research on ethnicity classification, economists and historians have long examined occupations among German-born, Polish-born, Russian-born, and Jewish immigrants, and indeed found huge heterogeneity in occupations by origin. Earlier works point out that the development of the agricultural economy in the U.S. was highly related to immigrants from Russia (Saloutos, 1976; Bodnar, 1976), Poland (Thomas and Znaniecki, 1996), and Germany (Luebke, 1990), especially in the Midwest where the climatic condition was similar to that in Central and Eastern Europe (Steckel, 1983). In contrast, Russian-born Jews were more likely to reside in the East Coast and California and were much less likely to work in agriculture (Simon, 1997). During the period of rapid urbanization in the early 20th century U.S., there were an increasing number of new jobs appearing in cities following economic growth at the local level (Kim, 1998), and Russians and Poles were involved in the expansion of the manufacturing sector in major Midwest industrial cities, such as Chicago (Pacyga, 2003) and Detroit (Zunz, 2000). While many Russian-born Jews also worked in manufacturing, they were still more likely to work in the business sector (e.g., Simon, 1997; Polland and Soyer, 2013).

Finally, a key question regarding ethnic occupational patterns is: what was the role of self-selection on migration in shaping occupational segregation by ethnicity among Russian-born immigrants in the U.S.? In theory, selection should lower the degree of ethnic occupational segregation among Russian-born immigrants in the U.S. compared to that in Russia, and one reason is that ethnic residential patterns in Russia were related to both selection on migration and occupational segregation. Specifically, three ethnic minority groups mainly resided in western Russia (see Figure 1) and in cities (e.g., Simon, 1997; Wetherell and Plakans, 1999; Snyder, 2006), which were already areas "in favor of emigration"[6] with more urban-type jobs (e.g., agriculture should not be the main sector). On the other hand, as Russian-ethnic immigrants were also more likely to be originally from these areas, ethnic differences in place of origin should be smaller among Russian-born immigrants in the U.S. than Russian citizens who stayed behind. In Section 4, we will empirically test whether ethnic occupational segregation still existed among Russian-born immigrants in the U.S. and present a case study of manufacturing jobs that compares occupational segregation among Russians in the U.S. and Russia.

3 Ethnicity Classification: Data and Methods

3.1 Classification Algorithm

The basic idea of the ethnicity classification algorithm is to use information on name and mother tongue—which are widely used to identify ethnicity in anthropology (e.g., Guglielmino et al., 2000; Duranti, 2007), as discussed in Section 2.1—to classify ethnicity among Russian-born immigrants. Specifically, we design a scoring system and assign a score for each of the four major ethnic groups, namely, Russian, Jewish, German, and Polish ethnics. For individual i, the score S_j^i for ethnic group j is calculated as:

$$S_j^i = f(\alpha_0 + \mathbf{X}_i\alpha_1 + \mathbf{Y}_i\alpha_2 + \mathbf{Z}_i\alpha_3) \tag{1}$$

$f(\cdot)$ calculates a score of ethnicity for i based on i's characteristics of mother tongue

[6] As discussed earlier, for example, more than 80% of Jews lived in cities in Russia and worked in either the manufacturing (secondary) and services (tertiary) sector (Simon, 1997). Similar geographic and occupational patterns also existed for Polish and German ethnics in Russia (e.g., Snyder, 2006; Sammartino, 2010).

($\mathbf{X}_i$), first name ($\mathbf{Y}_i$), and last name ($\mathbf{Z}_i$). We classify ethnicity based on $\hat{e}_i = \mathrm{argmax}_j\, S^i_j$. Here, $\mathbf{X}_i$ is a vector of mother tongue dummies (Russian, Yiddish/Hebrew, German, Polish), and $\mathbf{Y}_i$ and $\mathbf{Z}_i$ are vectors that measure probabilities that i's first and last name belong to specific linguistic origins.

It is easy to create $\mathbf{X}_i$ (language-based ethnicity), as the U.S. census surveys mother tongue. To construct $\mathbf{Y}_i$ and $\mathbf{Z}_i$ (name-based ethnicity), we employ a Bayes classifier, in which *training data* are ethnicity casebooks that are name dictionaries of Russian (Unbegaun, 1972), Jewish (Stern and Rottenberg, 1998), German (German Research Association, 1990), and Polish (Hoffman, 2001) ethnics.[7] These casebooks provide information on etymology and cultural-linguistic origin of names, and individuals' name-based ethnicity and their languages (and thus language-based ethnicity) can be determined.[8]

Specifically, to use the Bayes classifier, we first decompose names of Russian-born immigrants in census data into three- and four-character strings.[9] For each string, we first count the number of times that the string appears in four ethnicity casebooks, and then probabilistically calculate name-based ethnicity based on the following equation:

$$E^i_j = \sum_{k=1}^{n_i} P(L_k)P(e^k_j|L_k) \tag{2}$$

where $\{L_k\}$ $(k = 1, 2, \cdots, n_i)$ is the set of three-and four-character strings within i's name. $P(L_k)$ is the probability that the string L_k appears in all casebooks, and $P(e^k_j|L_k)$ is the probability that L_k belongs the j-th ethnic origin (e_j). Thus, the linguistic origin of i's

[7] In addition, we can conduct analyses based on (a) Russian census records (Central Statistical Committee of Russia, 1897), and (b) online training data (e.g., Wikipedia), which follows the norm in computer science (Treeratpituk and Giles, 2012). Empirical results based on different types of training data are very similar.

[8] Note that these casebooks covers all the historical regions of the specific country, but on the other hand, does not include names of recent immigrants. For example, the German name casebook includes German ethnic names originally in East Prussia, now in Poland, Lithuania, and Russia; however, it does not include names of Turkish and Balkan immigrants that came to Germany in recent decades and added to the diversity of German names. This actually makes these ethnic name casebooks a better training dataset in this paper's context, as recent immigrants had not arrived in these countries back in the late 19th and early 20th century and thus cannot reflect the linguistic and cultural traditions.

[9] For example, for the name *KAHN*, we decompose it into the following strings: *$KA*, *KAH*, *AHN*, *HN$*, *$KAH*, *AHN*, *HN$*, where $ represents the first or last character of the name.

name is $\tilde{e}_i = \text{argmax}_j\, E^i_j$, and $\mathbf{Y}_i$ and $\mathbf{Z}_i$ contain dummies of linguistic origins of names.

The basic idea of name-based classification is that names could contain ethnic-specific linguistic elements. Many prior studies follow the same idea but only conduct deterministic or quasi-deterministic name matching for classification (e.g., Mateos, 2007; Foley and Kerr, 2013; Ghani et al., 2014) in a deterministic or quasi-deterministic[10] manner, which is arguably valid if data are of high data quality (e.g., patent database in Foley and Kerr, 2013). In this paper, however, census records are of relatively low quality for two reasons. First, the digitization of census data is subject to transcription errors (Ruggles et al., 2019), such as typographic errors and misspellings.[11] Second, immigrants might change names after arrival. Although only a small proportion of immigrants completely Americanized their names—especially last names—in the early 20th century U.S. (Biavaschi et al., 2017), partial Americanization still makes it impossible to deterministically classify name-based ethnicity.[12] That said, it is likely to probabilistically classify ethnicity based on parts of the names that reflect idiosyncrasies of specific languages.[13]

After obtaining $\mathbf{X}_i$, $\mathbf{Y}_i$, and $\mathbf{Z}_i$, we revisit Equation 1 and determine "overall ethnicity" based on language-based and name-based ethnicity. The parameters $\{\alpha_j\}$ are trained through a classifier $f(\cdot)$ based on training data, i.e., ethnicity casebooks that now contain actual ethnicity and two types of calculated ethnicity ($\mathbf{X}_i$, $\mathbf{Y}_i$, and $\mathbf{Z}_i$). Specifically, in each casebook, we set $S^i_j = 1$ if e is i's actual ethnicity, and 0 otherwise. For example, for a Russian-born Jew *Leon Trotsky*, the score $s_{e=J} = 1$, and $s_{e=R} = s_{e=G} = s_{e=P} = 0$. We mainly use OLS as $f(\cdot)$ as it is computationally friendly and easy to interpret, but empirical

[10] It is still possible to assign probabilities for name-ethnicity pairs even in the deterministic algorithm if a name exists in multiple ethnic groups. For example, if there are 4 records of *Johnson* (say, 3 British and 1 Swedish) in training data, then the probability that Johnson is of British ethnicity is 75%, and the probability that Johnson is of Swedish ethnicity is 25%.

[11] For example, for an immigrant whose surname is *Eisenhauer*: while *Eisenhauer* is found in the German ethnic dictionary and thus of German name-based ethnicity, the name might be misspelled as *Eisenhouer* (because of misreading, mispronunciation, and blurred digitization), and cannot be found in any dictionaries.

[12] For example, *Eisenhauer* might be changed to *Eisenhower*, which cannot be found in any dictionaries.

[13] For example, while both *Eisenhouer* (transcription errors) and *Eisenhower* (name Americanization) cannot be found in German name dictionaries, it is still possible to probabilistically identify his German name-based ethnicity because of the typical German linguistic element *Eise*.

results based on other classifiers (e.g., logit, probit, SVM) are very similar, and in Section 3.3 and Section 4 we only report results of OLS-based classification.

3.2 Performance in Test Data

We test the performance of our algorithm in test data retrieved from different sources. The first dataset is constructed based on a series of records in Imperial Russia, including Russian census data and the Jewish families census records (Central Statistical Committee of Russia, 1897), Baptism records (Ancestry.com, 2014a), the database of East Prussians from Russia (Anuta, 2002), and Roman Catholic Church Books Index (Ancestry.com, 2014b). The second dataset consists of Russian politicians between 16th and 20th century. The first dataset contains 1,200 people that are randomly chosen from the population data. The second dataset contains 200 people. We collect sociodemographic characteristics for people in our data and are thus able to confirm their ethnicity.[14] Each ethnic group has 300 persons in the first dataset and 50 persons in the second dataset.

We use four classical measures in the machine learning literature to evaluate our method: precision, recall, F-measure, and accuracy. For an ethnic group e, these four measures focus on four classification outcomes: true positive (*tp*): a person of ethnicity e is correctly classified as in group e; true negative (*tn*): a person *not* of ethnicity e is correctly classified as *not* in group e; false positive (*fp*): a person *not* of ethnicity e is incorrectly classified as in group e; false negative (*fn*): a person of ethnicity e is incorrectly classified as *not* in group e. Specifically, *precision*, *recall*, and F are defined by:

$$Precision_j = \frac{tp}{tp + fp}, Recall_j = \frac{tp}{tp + fn}, F_j = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

[14] For the first dataset: by definition, individuals retrieved from Jewish families records in the Russian census, East Prussians records, and Polish Roman Catholic Church Books Index are of Jewish, German, and Polish ethnicity, respectively. Individuals who were documented in the Orthodox Church are identified as of Russian ethnicity. For the second dataset: the Russian politicians' biographic information (mother tongue, place of birth, family ancestry, ethnicity) are publicly available.

and *accuracy* is the overall fraction of cases that are classified correctly.

[Insert Table 3 here.]

Table 3 presents results. Panel A focuses on data on Russian citizens in historical census and other population records. Panel B focuses on data on Russian politicians. Our method is robust in both tests: the false positives and negatives are rare, and overall accuracy rates are very high. Our method performs better in the second test dataset. Note that, however, a test on Russian politicians is "biased" as famous people's names are more likely to be collected in name dictionaries, and the sample size in Panel B is smaller. Hence, the classification accuracy is upward biased.

3.3 Classification Results in 1930 Census Data

We close this section by reporting results of ethnicity classification among Russian-born immigrants in IPUMS' restricted version of digitized 1930 U.S. census data (Ruggles et al., 2019) that contain publicly available demographic and socioeconomic variables, as well as name records. We only classify men's ethnicity because (a) many women changed their last name after marriage, and it is thus impossible to identify their name-based ethnicity, and (b) for the substantive research question of this paper, unlike men in the U.S., many women did not participate in the labor force (Boustan et al., 2014).

[Insert Table 4 here.]

Table 4 presents results of ethnicity classification among 609,200 Russian-born male immigrants in the 1930 U.S. census. Panel A shows that over 60% of Russian-born immigrants were Jews. While this looks surprising in the first place, it is actually consistent with historical findings that Jews had stronger incentives to migrate (e.g., Nathans, 2002, 2006), and Jews were indeed the majority ethnic group among Russian-born immigrants in the U.S. (Simon, 1997). This number is also consistent with mother tongue records of the 1930 U.S. census that over 60% of Russian-born immigrants spoke Yiddish or Hebrew, two ethnic languages that were rarely spoken by non-Jewish groups. Besides, 25.8% of Russian-born immigrants were of German ethnicity, 8.4% were of Russian ethnicity, and 2.7% were of Polish ethnicity.

We further examine the interaction between ethnicity and years of arrival. Migration from Russia to the U.S. started to rise around 1900; over 50% of immigrants in the sample arrived in the U.S. between 1905 and 1914. On the other hand, very few immigrants arrived during 1915-1919 (World War I) and 1925-1930 (after the passage of immigration restriction laws, see Ngai, 1999). We observe the similar trends in migration in each ethnic group. The major differences are that Russian-born German ethnics generally arrived in the U.S. earlier, and Russian and Polish ethnics arrived later.

Finally, we describe geographic distributions of four ethnic groups by county in Figure 2. The color in each county reflects the number of Russian-born immigrants of specific ethnicity. These maps show that for all ethnic groups, immigrants were primarily concentrated in the East Coast (especially Greater New York and Boston area), Chicago, and Southern California. Jews were spatially more concentrated than other groups. A large proportion of German ethnics resided in the Midwest, Mountain states, and Northwest; in contrast, very few Russian-born immigrants of other ethnicity lived in these areas.

[Insert Figure 2 here.]

4 Occupational Segregation: Empirical Analysis

4.1 Summary Statistics: Immigrants and Industries in 1930

In Table 5, we begin with summary statistics of Russian immigrants' basic demographic and socioeconomic characteristics (Panel A) and labor market characteristics (Panel B). In 1930, four ethnic groups had the similar average age—approximately 42 years old—although German ethnics were slightly older. Overall, immigrants were much older than natives due to the lack of new immigrants following immigration restriction laws (Ngai, 1999). Four ethnic groups from Russia (as well as other immigrants) had similar years since migration. Among all Russian-born immigrants, Jews were substantially more likely to live in urban areas: in fact, only less than 2% of Russian-born Jews lived in rural areas.[15] On the other hand, rates of urban residence were lower in other groups, especially among German ethnics (71.8%). But in general, immigrants were far more likely to live in urban areas than natives. Jews and German ethnics had slightly larger households, were more likely to be U.S. citizens, and had higher literacy rates. The marriage rate ranged from 70% to 80% across groups. Finally, most Russian-born immigrants, regardless of ethnicity, could speak English.

Panel A of Table 5 also shows that Russian-born immigrants had significantly different sociodemographic characteristics with other immigrants in the U.S. Russian-born immigrants were more likely to live in cities and were more likely to be citizens and be married. Russian-born immigrants were also more likely to be literate and had better English skills. In addition, there were significant within-population differences in these individual characteristics among Russian-born immigrants: while Panel A shows some insignificant pairwise differences (e.g., average age among Jews and Polish ethnics), but jointly, four ethnic groups' sociodemographic characteristics were significantly different.

[Insert Table 5 here.]

We further study labor market characteristics in Panel B of Table 5. Russian-born immigrants had a higher literacy rate and were more likely to speak English well (see Panel A), which could explain that they had a higher employment rate, higher occupation-based earnings,[16] and were more likely to work in higher-paying industries, including fi-

[15] *Urban* was defined by the Census Bureau. In 1930, the Census Bureau largely maintained the definition of *urban* in the 1920 census: *urban areas* were considered to be cities and incorporated places of 2,500 inhabitants or more; the extension of the definition of *urban* in the 1930 census was that townships and other political subdivisions (not incorporated as municipalities) having a total population of 10,000 or more, and a population density of 1,000 or more per square mile, were also considered to be urban (Ruggles et al., 2019).

[16] The variable *occupational scores* is the occupation-based earnings measured by the median total income (in hundreds of 1950 dollars) of all persons with that occupation in 1950 (Ruggles et al., 2019). The 1930 U.S. census (and all previous censuses) did not survey individual income. A widely used proxy for income

nance/business, professional services, and trade, which had higher average occupation-based earnings in the U.S. Within the Russian-born immigrants, labor force participation rates were around 90% in all four ethnic groups and were insignificantly different by ethnicity. Conditional on labor force participation, Polish ethnics had a slightly lower employment rate. On average, Jews had the highest occupation-based earnings among Russian-born immigrants, while Polish ethnics had the lowest occupation-based earnings. We finally study Russian-born immigrants' propensity to work in higher paying industries. Again, Jews were substantially more likely to work in higher paying industries, while Polish ethnics were less likely to work in such industries.

Note that Russian-born Jews' occupation-based earnings exceed earnings of both natives and other immigrants, which is consistent with findings of Abramitzky et al. (2014): during the age of mass migration, immigrants did *not* face the earnings penalty upon arrival, and did not economically assimilate; while the economic condition of the country of origin was a good predictor of individuals' post-migration outcomes, an exception was the Russian-born immigrant population who had relatively high occupation-based earnings even though Imperial Russia and its successors (e.g., Russian Republic, and then USSR) were less developed. Our analysis in Section 3.3 and 4.1 is in line with their findings.

Before moving to the analysis of occupational segregation, we present a brief overview of industries in Table 6. We adopt the system of occupation and industry classification made by the Census Bureau (1950), which is digitized by IPUMS (Ruggles et al., 2019).[17] Ten industries are listed in Table 6. The first four columns show that natives and immigrants (Russian-born immigrants in particular) were concentrated in significantly different industries. Compared to natives, Russian-born immigrants were more likely to work in construction, manufacturing, personal/entertainment services, and trade. On the other hand, Russian-born immigrants were less likely to work in public administration,

in the prior literature is IPUMS' occupational score variable constructed based on available individual-level income data in 1950 and the 1950 industry classification system (Census Bureau, 1950).

[17] In 1930 census data, approximately 15% of people had occupations that were unclassified in the industry classification system (Census Bureau, 1950), such as students and retired workers.

agriculture, mining, and transportation/communication. These industries also had different levels of occupation-based earnings. Note that although natives had relatively high occupation-based earnings in transportation/communication (see row 10), immigrants (including Russian-born immigrants) earned much less in this industry, and in this paper we do not consider it as a higher paying industry in the subsequent analysis.

[Insert Table 6 here.]

4.2 Occupational Segregation among Russian-Born Immigrants

We now present our main results of occupational segregation by ethnicity in Table 7. In Section 4.1, we show that occupational segregation existed among immigrants and natives in 1930. Focusing specifically on Russian-born immigrants, we observe high degrees of within-population occupational segregation by ethnicity as well. German ethnics had the highest proportion (24.94%) of people working in agriculture; while very few Jews (0.74%) worked in this industry. Polish ethnics had the highest proportion (7.95%) of people working in mining; again, very few Jews (0.06%) worked in this industry. Jews were also less likely to work in transportation and communication, a major industry in the then U.S. On the other hand, almost half of all Jews worked in wholesale and retail trade, although other ethnic groups also had significant proportions of people working in this industry.

[Insert Table 7 here.]

Four ethnic groups had similar proportions of people in other industries. In all ethnic groups, very few people worked in public administration. Jews were slightly more likely to work in construction, finance/business, and personal/entertainment services, and Russian ethnics were slightly more likely to worked in professional services, but the differences were generally small. Manufacturing was a major industry across groups, while Polish ethnics were substantially more likely to work in this industry.

The above findings of occupational segregation explain Russian-born immigrants' economic performance in the early 20th century U.S. A further question is: did Russian-born Jews' advantages in occupational outcomes still exist after controlling for covariates? Table 8 explores this question in a regression framework. In the first four columns, we run OLS regressions of occupational scores on ethnicity dummies as well as other individual characteristics, state fixed effects, and industry fixed effects, with Russian ethnics (the main ethnic group back in Russia) as the reference group. We cluster standard errors at the state level.[18] Column 1 shows the baseline estimation: Jews' annual occupation-based earnings were 378 dollars (in 1950) higher than Russian ethnics, while Polish ethnics' annual occupation-based earnings were 279 dollars lower than Russian ethnics.[19] There were no significant differences in occupation-based earning between Russian ethnics and German ethnics. This pattern becomes weaker when we add individual characteristics and state controls. Moreover, ethnic differences in occupation-based earnings appear to be even smaller after including industry fixed effects (column 4): Jews' annual occupation-based earnings were only 80 dollars (in 1950) higher than Russian ethnics, and Polish ethnics' annual occupation-based earnings were only 107 dollars lower than Russian ethnics. This is consistent with descriptive findings in the last four columns of Table 7 that show that Jews had relatively higher occupation-based earnings in almost all industries, while Polish ethnics had lower within-industry occupation-based earnings, but such differences in earnings by ethnicity within industry were significantly smaller than differences in earnings by ethnicity across industries. An explanation is that in each industry, many jobs have fairly standard ranges of requirements that are similar within this industry, which include human capital characteristics, skills, age and gender, and other characteristics, and thus the degree of within-industry occupational segregation by ethnicity appears to be lower.

[Insert Table 8 here.]

[18] Clustering standard errors at other levels (such as county, enumeration district, and industry) does not change the empirical conclusion.

[19] As a comparison, the mean household income in 1950 was 3,300 dollars (in 1950). This suggests the magnitudes of the differences reported in Column 1 are fairly large.

We also aggregate three higher paying industries—finance/business, professional services, and trade—into one category and examine differences in the propensity to work in higher paying industries by ethnicity in Table 8. Results show that consistent with descriptive findings in Table 7, Jews were indeed more likely to work in higher paying industries, even after controlling for individual characteristics, state fixed effects, and excluding agriculture (see column 7). However, this was mainly because of Jews' concentration in trade: column 8 shows that Russian-born Jews were not significantly more likely to work in finance/business and professional services than other ethnic groups.

One question regarding ethnic occupational patterns is: to what extent the observed occupational segregation by ethnicity among Russian-born immigrants in the U.S. replicates the heterogeneity that existed in Russia? On one hand, ethnic occupational segregation among Russian-born immigrants in the U.S. should be substantially different from that in Russia's *entire labor market*, as Russian immigration was positively selected in terms of human capital characteristics and urban residence. For example, 1897 Russian census data (Central Statistical Committee of Russia, 1897) show that approximately 75% of Russian laborers worked in agriculture, with a disproportionately large share of Russian ethnics (as ethnic minorities were more likely to reside in cities), but Table 6 shows that only 8% of Russian-born immigrants (and 9% of Russian ethnics) worked in agriculture in the U.S., which could be explained by selection that farmers were less likely to migrate.

[Insert Table 9 here.]

On the other hand, ethnic occupational segregation among Russian-born immigrants in the U.S. appears to be more similar to that in Russia's major cities, as both types of observations partially account for selection on employment. Table 9 presents two case studies. In Panel A, we compare classes of manufacturing jobs—the majority industry among U.S. immigrants—in Russia's St. Petersburg (Bater, 1976; also see Table 2) and the U.S. (1930 census data, Ruggles et al., 2019) by ethnicity. Results show that patterns of ethnic occupational segregation tend to be similar in two countries, especially for managerial jobs and self-employment. In Panel B, we compare industries among Jewish workers in Russia (Simon, 1997; also see Table 2) and the U.S. (1930 census data, Ruggles et al., 2019). Results similarly show that Jewish concentrations in manufacturing and trade appear to be similar in both countries. The replication of ethnic occupational segregation could be because that the acquisition of urban-type jobs in both Russia (note that most Jews resided in cities in Russia) and the U.S. was selected based on human capital characteristics and residential choices. One caveat of the above results is that Russian data mainly focus on occupational segregation by Jewish and non-Jewish ethnicity (but not other ethnic groups). However, Table 9 still presents some empirical evidence of similarities in ethnic occupational segregation in Russia and the U.S.

4.3 Spatial Concentration of Immigrants and Occupational Outcomes

We finally turn to study the relationship between the spatial concentration of immigrants and labor market outcomes. Researchers have long discussed the economic impacts of ethnic enclave residence and find mixed results (Cutler et al., 2008). On one hand, leaving ethnic enclaves signals social assimilation (Bleakley and Chin, 2010; Xu, 2017), which is generally associated with economic status. On the other hand, immigrants are more likely to get support from social networks when living in ethnic enclaves (Munshi, 2003; Patacchini and Zenou, 2012), and thus researchers find ethnic enclave residence leads to better labor market outcomes after taking sorting into account (Edin et al., 2003).

In Table 10, we explore the relationship between the spatial concentration of immigrants and occupational outcomes by ethnicity. Specifically, we measure the spatial concentration of immigrants by ethnicity within the Russian-born immigrant population based on the size of the co-ethnic enclave (i.e., the number of immigrants of that specific ethnicity) at the county level and follow the idea of Munshi (2003) that counts old (thus more established) and new immigrants separately. We study three types of outcomes: employment status, occupational score (i.e., earnings), and the likelihood of working in a higher paying industry. Specifically, we estimate the following OLS model:

$$L_i = \alpha_0 + \alpha_1 N_{c(i)}^{e(i)} + \alpha_2 N_{c(i)}^{-e(i)} + \alpha_3 n_{c(i)}^{e(i)} + \alpha_4 n_{c(i)}^{-e(i)} + \mathbf{X}_i \alpha_5 + \varepsilon_i \tag{4}$$

where i indexes individual; $e(i)$ is i's ethnicity classified in Section 3.3, and $c(i)$ is i's county of residence. L_i is a specific type of labor market outcome. $N_{c(i)}^{e(i)}$ is the number of old Russian-born immigrants (who arrived before 1920) of $e(i)$ (i.e., i's co-ethnics) living in $c(i)$; $N_{c(i)}^{-e(i)}$ is the number of old Russian-born immigrants (who arrived before 1920) whose ethnicity is *not* $e(i)$ (i.e., i's compatriots of other ethnicity) living in $c(i)$. We similarly use $n_{c(i)}^{e(i)}$ and $n_{c(i)}^{-e(i)}$ to denote the number of of new Russian-born immigrants (who arrived in or after 1920). $\mathbf{X}_i$ is a vector of control variables introduced in Table 9. Note that as only one census year is used in the empirical analysis, time and cohort controls are not included in our specifications.

Many economists argue that Equation (1) has several types of econometric issues. First, there might be omitted variables not included in $\mathbf{X}_i$ that are correlated with both the spatial concentration of immigrants and labor market outcomes. Specifically, self-selection on migration by ethnicity might introduce bias into the regression's estimates. For example, ethnic enclave residence might be positively correlated with both emigration from Russia and the reliance on social networks in the U.S., through which an ethnic group achieved better labor market outcomes. Second, L_i might reversely affect residential choices. Finally, the spatial concentration of immigrants might be incorrectly measured, as census enumeration occasionally missed respondents (Ruggles et al., 2019) or recorded wrong information about respondents, especially for immigrants as some of them could not speak English well.

In the classical literature of labor economics (Card, 2001), one standard approach is to use historical settlements of immigrants as an instrumental variable (IV) for current settlements. This is based on the idea that immigrants prefer to reside in areas where historically the same group of immigrants resided (Bartel, 1989). Furthermore, to serve as a valid IV, historical settlements should not be correlated with any omitted variables in ε_i that also affect labor market outcomes. In many prior studies (e.g., Ottaviano and Peri, 2006; Saiz, 2007; Olney, 2013; Accetturo et al., 2014),[20] scholars argue that this assumption is satisfied because with a sufficiently long time lag, the local economic condition at the time of historical settlements should be unrelated to the current local economic condition.

[Insert Table 10 here.]

In this paper, we use historical settlements of Russian-born immigrants by ethnicity in 1880 to construct IVs for ethnic enclaves of old immigrants. To do so, we rerun our machine learning algorithm of ethnicity classification in the sample of Russian-born immigrants in 1880 full-count U.S. census data (Ruggles et al., 2019) and measure county-level ethnic enclaves. The IVs should be especially useful in this paper for two reasons: (a) the U.S. experienced massive transformation in the society and economy throughout the late 19th and early 20th century (e.g., Kim, 1998), and the local economic condition in 1880 was very different from that in 1930 given a half-century lag; (b) these IVs are especially effective if historical settlements of immigrants are measured based on the year when the immigrant population was small (Accetturo et al., 2014) as earliest immigrants should have no prior settlement patterns to follow,[21] which is exactly the case during the age of mass migration (Tabellini, 2019). Indeed, there were only about 32,000 Russian-born immigrants, regardless of ethnicity, who lived in the U.S. in 1880, and they were among the earliest immigrants from Imperial Russia in the late 19th century U.S.[22] Hence, Russian-born immigrants in 1880 were unlikely to be able to follow any prior settlement patterns. In the

[20] Starting from the paper by Card (2001), this approach has been widely used in the economic analysis of immigration. For example, using historical settlements of immigrants to instrument for immigrants' residential choices, Ottaviano and Peri (2006) find positive impacts of cultural diversity on natives' productivity; Olney (2013) shows that immigration leads to firm expansion; Saiz (2007) and Accetturo et al. (2014) show that immigration flows positively affect rents and housing values in both the U.S. and Europe; Tabellini (2019) finds that immigration flows affected natives' political attitudes during the age of mass migration in the U.S., the same historical period studied in our paper.

[21] In some extreme cases (e.g., Waldinger, 2017), scholars even believe that immigrants' (earliest) settlements in history were arguably random, and thus historical settlements created a natural experiment on immigrants' residential choices in terms of geographic characteristics in the host country.

[22] There were only about 5,000 and 700 Russian-born immigrants in the 1870 and 1860 U.S., respectively.

first two columns of Table 10, we run first-stage regressions of the number of residents of same and different ethnicity in 1930 on historical settlements in 1880. We find that historical settlements successfully predict ethnic enclave residence in 1930, and first-stage results present no evidence of weak instruments.

We first investigate employment status in column 3 of Table 10. Column 3 suggests that the number of co-ethnics in the county was positively associated with employment status, and similarly, the number of Russian-born immigrants of other ethnicity was also positively related to employment status, while new immigrants—regardless of ethnicity—were negatively correlated with employment status. Column 4 shows IV results, which suggest that the effect of the concentration of established co-ethnics appears to be underestimated in the OLS regression; on the other hand, the effect of the concentration of Russian-born immigrants in other ethnic groups becomes insignificant. We find that individuals' employment status was mainly affected by established co-ethnics; immigrants of other ethnicity—even if they were also born in Russia—had no effects on employment status.

We turn to focus on occupation-based earnings and the likelihood of working in higher paying industries in the next four columns. Again, we first present results of OLS regressions and then present IV estimates. Similar to our findings in columns 3 and 4, we observe that the spatial concentration of established co-ethnics—but not compatriots of different ethnicity—was positively associated with both earnings and the likelihood of working in a higher paying industry. These results are consistent with the classical conclusion in labor economics that immigrants get support from ethnic enclaves in the labor market (Edin et al., 2003; Patacchini and Zenou, 2012). More importantly, we highlight the importance of having an accurate measure of ethnicity (Munshi, 2003): measuring immigrant enclaves using a coarse measure of ethnicity might involve immigrants who do not contribute to ethnic social networks, further leading to underestimation of network effects.

In Table 11, we run three additional regressions of labor market outcomes on the size of "compatriot networks" measured by the number of Russian-born immigrants at the county level, regardless of ethnicity. Results of Table 11 are qualitatively similar to earlier findings. However, the effect size becomes significantly smaller in each regression, compared with Table 10. Table 11 again highlights the importance of taking ethnic heterogeneity into consideration by presenting that Russian-born immigrants mainly relied on social networks defined by their specific ethnicity, and other ethnic groups born in Russia had weaker, if not null, effects.

[Insert Table 11 here.]

We conclude the empirical section by discussing heterogeneous effects of ethnic networks in different ethnic groups in Table 12. We measure the effects in terms of three types of labor market outcomes: employment status, occupation-based earnings, and the likelihood of working in a higher-paying industry. Results of the table show that effects of ethnic social networks appear to be strongest among German ethnics. This could be because of German ethnics' concentration in agriculture, in which social networks should play a crucial role in communication, marketing, and collaboration on agricultural technologies. However, column 11 shows that German ethnic networks also affected occupational prestige (measured by whether an immigrant entered a higher paying industry, which does not include agriculture). Besides, Jewish networks also had statistically significant effects on all three types of labor market outcomes with smaller effect sizes. Finally, we find that Russian ethnic networks significantly affected employment status and occupational prestige, and Polish ethnic networks significantly affected earnings and occupational prestige. In general, Table 12 presents similar qualitative patterns across four ethnic groups that co-ethnic networks significantly influenced the labor market outcomes.

[Insert Table 12 here.]

5 Conclusion

Scholars have long discussed occupational segregation between immigrants and natives (Borjas, 1986), and across immigrant populations (Fairlie and Meyer, 1995). A traditional way to classify immigrant populations is based on the country of birth, which neglects ethnic heterogeneity within a co-birth immigrant population (Xu, 2019). This points out possible occupational segregation by ethnicity within an immigrant population because different ethnic groups might have different social, economic, and cultural backgrounds.

In this paper, we present a case study of occupational segregation by ethnicity among Russian-born immigrants in the age of mass migration in the U.S. Russia was a major sending country of U.S. immigrants in the early 20th century and had a high degree of ethnic diversity. In the first part of the paper, we develop a machine learning method to classify ethnicity using Russian immigrants' name and mother tongue information surveyed in the 1930 U.S. census. Both name-based ethnicity (e.g., Chibnik, 1991; Mateos, 2007) and language-based ethnicity (Duranti, 1991) are widely used to measure immigrants' ethnicity in anthropology and have been applied in economics (e.g., Foley and Kerr, 2013; Zhang, 2016). In particular, our method allows fuzzy classification, which is particularly useful in the context of this paper that individuals' digitized census records are of low data quality because of, e.g., transcription errors and name changes.

Based on this constructed ethnicity variable, we indeed find a high degree of occupational segregation by ethnicity within the Russian-born immigrant population. Specifically, Russian-born Jews were significantly more likely to work in trade and were less likely to work in agriculture and mining. German ethnics were particularly engaged in agriculture, and Polish ethnics were concentrated in manufacturing. The overall pattern of occupational segregation remains unchanged after we include individual characteristics, geographic controls, and industry controls as covariates within a regression framework, and is consistent with earlier qualitative and descriptive findings in historical research (e.g., Luebke, 1990; Thomas and Znaniecki, 1996; Simon, 1997; Haines, 2000; Zunz, 2000).

We further show evidence of the relationship between the spatial concentration of immigrants and labor market outcomes. We find that the size of the ethnic enclave—especially the enclave of more established immigrants—was positively related to employment status, occupation-based earnings, and the likelihood of working in a higher paying industry. This is not surprising and has been pointed out by a large body of literature in network economics (e.g., Edin et al., 2003; Cutler et al., 2008; Patacchini and Zenou, 2012); what is more interesting and policy relevant is that the number of established co-ethnic compatriots mattered most in the labor market, while compatriots of other ethnicity—even if they were also born in Russia—had weaker or null effects on labor market outcomes.

In this paper, the methodological analysis on ethnicity classification presents a finer measure of immigrant origin and leads to new findings of occupational segregation. While this paper focuses on the history of immigration, it still has important implications for immigration issues in contemporary contexts. For example, the U.S. Census Bureau broadly defines "Hispanic immigrants" as those originally from Latin America. While Hispanic immigrants do speak the same mother tongue, this coarse measure neglects huge heterogeneity in social organizations (Portes, 1987), labor market outcomes (Borjas, 1982), interactions with natives (South et al., 2006), marriage patterns (Gilbertson et al., 1996), and public health outcomes (e.g., Hummer et al., 2000; Acevedo-Garcia et al., 2007) within the Hispanic population. Moreover, many major sending countries of immigrants in Latin America (e.g., Mexico and Peru) as well as other regions of the world (e.g., India) have high degrees of ethnic diversity, and it is possible that immigrants of different ethnic origins from these countries form different ethnic social networks even within the co-birth population after their arrival. For example, Munshi (2003) finds within-population differences in occupational choices among Mexican immigrants, and such differences are strengthened by the social networks defined based on specific community of origin, rather than a coarse measure of country-based origin. Another similar example in an European context involves immigration after the Yugoslav Wars to Germany in the 1990s (Carter, 1993; Grün, 2009), when many refugees had similar biological ancestries and the same nationality even after the breakup of Yugoslavia but still considered themselves to belong to different ethnic groups (e.g., Bosnian, Croatian, and Serb ethnics originally from Bosnia and Herzegovina) due to political, religious, and social reasons. These studies—including this paper—point out the necessity to have a fine measure of ethnicity in order to accurately evaluate the role of ethnicity and ethnic social networks in affecting immigrants' labor market outcomes, cultural assimilation, social behaviors, and long-term health status.

Ethical Statements

The authors received no funding for this study. The authors declare that they have no conflict of interest.

References

[1] 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.

[2] Accetturo, Antonio, Francesco Manaresi, Sauro Mocetti, and Elisabetta Olivieri. 2014. "Don't Stand So Close to Me: The Urban Impact of Immigration." *Regional Science and Urban Economics*, 45, 45 - 56.

[3] Acevedo-Garcia, Dolores, Mah-J. Soobader, and Lisa F. Berkman. 2007. "Low Birthweight among US Hispanic/Latino Subgroups: The Effect of Maternal Foreign-Born Status and Education." *Social Science & Medicine*, 65(12), 2503 - 2516.

[4] Ancestry.com. 2014a. *Russia, Select Births and Baptisms, 1755-1917* [database on-line]. Provo: Ancestry.com.

[5] Ancestry.com. 2014b. *Poland, Roman Catholic Church Books Index, 1742-1964* [database on-line]. Provo: Ancestry.com.

[6] Anuta, Michael J. 2002. *East Prussians from Russia*. Baltimore: Genealogical Publishing Co.

[7] Bandiera, Oriana, Imran Rasul, and Martina Viarengo. 2013. "The Making of Modern America: Migratory Flows in the Age of Mass Migration." *Journal of Development Economics*, 102, 23 - 47.

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

[9] Bater, James H. 1976. *St Petersburg: Industrialization and Change*. London: Edward Arnold.

[10] Blau, Francine D., and Larry M. Kahn. 2015. "Substitution between Individual and Source Country Characteristics: Social Capital, Culture, and US Labor Market Outcomes among Immigrant Women." *Journal of Human Capital*, 9(4), 439 - 482.

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

[12] Bleakley, Hoyt, and Aimie Chin. 2010. "Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants." *American Economic Journal: Applied Economics*, 2(1), 165 - 192.

[13] Bodnar, John. 1976. “Immigration and Modernization: The Case of Slavic Peasants in Industrial America.” *Journal of Social History*, 10(1), 44 - 71.

[14] Borjas, George J. 1982. “The Earnings of Male Hispanic Immigrants in the United States.” *Industrial and Labor Relations Review*, 35(3), 343 - 353.

[15] Borjas, George J. 1986. “The Self-Employment Experience of Immigrants.” *Journal of Human Resources*, 21(4), 485 - 506.

[16] Borjas, George J. 1994. “Immigrant Skills and Ethnic Spillovers.” *Journal of Population Economics*, 7(2), 99 - 118.

[17] Botticini, Maristella, and Zvi Eckstein. 2005. “Jewish Occupational Selection: Education, Restrictions, or Minorities?” *Journal of Economic History*, 65(4), 922 - 948.

[18] 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.

[19] Boustan, Leah Platt, Carola Frydman, and Robert A. Margo. 2014. *Human Capital in History: The American Record*. Chicago: University of Chicago Press.

[20] Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review*, 94(4), 991 - 1013.

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

[22] Card, David. 2001. “Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration.” *Journal of Labor Economics*, 19(1), 22 - 64.

[23] Carlton, Dennis W., and Avi Weiss. 2001. “The Economics of Religion, Jewish Survival, and Jewish Attitudes Toward Competition in Torah Education.” *Journal of Legal Studies*, 30(1), 251 - 275.

[24] Carter, F. W. 1993. “Ethnicity as a Cause of Migration in Eastern Europe.” *GeoJournal*, 30(3), 241 - 248.

[25] Census Bureau, 1950. *Alphabetic Index of Occupations and Industries: 1950*. Washington, D.C.

[26] Chernina, Eugenia, Paul Castaneda Dower, and Andrei Markevich. 2014. “Property Rights, Land Liquidity and Internal Migration.” *Journal of Development Economics*, 110, 191 - 215.

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

[28] Connor, Walker. 1993. *Ethnonationalism: The Quest for Understanding*. Princeton: Princeton University Press.

[29] Constant, Amelia, Liliya Gataullina, and Klaus F. Zimmermann. 2008. “Ethnosizing Immigrants.” *Journal of Economic Behavior & Organization*, 69, 274 - 287.

[30] Constant, Amelia, and Klaus F. Zimmermann. 2008. “Measuring Ethnic Identity and Its Impact on Economic Behavior.” *Journal of the European Economic Association*, 6(2-3), 424 - 433.

[31] Constant, Amelia, and Klaus F. Zimmermann. 2009. “Work and Money: Payoffs by Ethnic Identity and Gender.” *Research in Labor Economics*, 29, 3 - 30.

[32] Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 2008. “When Are Ghettos Bad? Lessons from Immigrant Segregation in the United States.” *Journal of Urban Economics*, 63(3), 759 - 774.

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

[34] Dowler, Wayne. 2000. *Classroom and Empire: The Politics of Schooling Russia’s Eastern Nationalities, 1860-1917*. Montreal: McGill-Queen’s University Press.

[35] Drydakis, Nick. 2011. "Ethnic Discrimination in the Greek Housing Market." *Journal of Population Economics*, 22(4), 1235 - 1255.

[36] Duranti, Alessandro. 1997. *Linguistic Anthropology*. Cambridge: Cambridge University Press.

[37] Eberhardt, Piotr. 2002. *Ethnic Groups and Population Changes in Twentieth Century Eastern Europe: History, Data and Analysis: History, Data and Analysis*. Abington: Routledge.

[38] 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.

[39] Eriksen, Thomas Hylland. 2002. *Ethnicity and Nationalism: Anthropological Perspectives*. London: Pluto Press.

[40] 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.

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

[42] German Research Association. 1990. *German Research Association Surname Book*. San Diego: German Research Association.

[43] Gilbertson, Greta A., Joseph P. Fitzpatrick, and Lijun Yang. 1996. "Hispanic Intermarriage in New York City: New Evidence from 1991." *International Migration Review*, 30(2), 445 - 459.

[44] Grün, Sonja. 2009. "Construction of Ethnic Belonging in the Context of Former Yugoslavia—The Case of a Migrant from Bosnia-Herzegovina." *Forum: Qualitative Social Research*, 10(3), Art. 22.

[45] 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.

[46] Haines, Michael R. 2000. "The Population of the United States, 1790 - 1920." In *The Cambridge Economic History of the United States: Volume 2*, eds., Stanley L. Engerman and Robert E. Gallman. New York: Cambridge University Press.

[47] Hirsch, Boris, Eke J.Jahn, Ott Toomet, and Daniela Hochfellner. 2014. "Do Better Pre-Migration Skills Accelerate Immigrants' Wage Assimilation?." *Labour Economics*, 30, 212 - 222.

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

[49] Hummer, Robert A., Richard G. Rogers, Sarit H. Amir, Douglas Forbes, and W. Parker Frisbie. 2014. "Adult Mortality Differentials among Hispanic Subgroups and Non-Hispanic Whites." *Social Science Quarterly*, 81(1), 459 - 476.

[50] Hunt, Edwin S., and James Murray. 1999. *A History of Business in Medieval Europe, 1200-1550*. Cambridge, U.K.: Cambridge University Press.

[51] Kalfa, Eleni, and Matloob Piracha. 2018. "Social Networks and the Labour Market Mismatch." *Journal of Population Economics*, 31(3), 877 - 914.

[52] Kalnins, Arturs, and Wilbur Chung. 2006. "Social Capital, Geography, and Survival: Gujarati Immigrant Entrepreneurs in the U.S. Lodging Industry." *Management Science*, 52(2), 233 - 247.

[53] Kim, Sukkoo. 1998. "Economic Integration and Convergence: U.S. Regions, 1840-1987." *Journal of Economic History*, 58(3), 659 - 683.

[54] Klier, John Doyle. 1995. *Imperial Russia's Jewish Question, 1855-1881*. New York: Cambridge University Press.

[55] Lieven, Dominic. 2006. *The Cambridge History of Russia, Volume 2: Imperial Russia, 1689–1917*. Cambridge: Cambridge University Press.

[56] Luebke, Frederick C. 1990. *Germans in the New World: Essays in the History of Immigration*. Champaign: University of Illinois Press.

[57] 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.

[58] Mateos, Pablo. 2014. *Names, Ethnicity and Populations: Tracing Identity in Space*. Berlin: Springer.

[59] Morton, Nicholas. 2009. *The Teutonic Knights in the Holy Land, 1190–1291*. Suffolk, U.K.: Boydell & Brewer.

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

[61] Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. “German Jewish Émigrés and US Invention.” *American Economic Review*, 104(10), 3222 - 3255.

[62] Mountain, Joanna L., and Neil Risch. 2004. “Assessing Genetic Contributions to Phenotypic Differences among ’Racial’ and ’Ethnic’ Groups.” *Nature Genetics*, 36, S48 - S53.

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

[64] Naimark, Norman M. 2002. *Fires of Hatred: Ethnic Cleansing in Twentieth-Century Europe*. Cambridge: Harvard University Press.

[65] Nathans, Benjamin. 2002. *Beyond the Pale: The Jewish Encounter with Late Imperial Russia*. Berkeley: University of California Press.

[66] Nathans, Benjamin. 2006. “The Jews.” In *The Cambridge History of Russia: Volume 2*, ed., Dominic Lieven. New York: Cambridge University Press.

[67] Ngai, Mae M. 1999. “The Architecture of Race in American Immigration Law: A Reexamination of the Immigration Act of 1924.” *Journal of American History*, 86(1), 67 – 92.

[68] O’Keefe, Siobhan, and Sarah Quincy. 2018. “Old Immigrants, New Niches: Russian Jewish Agricultural Colonies and Native Workers in Southern New Jersey, 1880–1910.” *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 4(1), 21 - 38.

[69] Olney, William. 2013. “Immigration and Firm Expansion.” *Journal of Regional Science*, 53(1), 142 - 157.

[70] Oreopoulos, Philip. 2011. “Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes.” *American Economic Journal: Economic Policy*, 3(4), 148 - 171.

[71] Ottaviano, Gianmarco I.P., and Giovanni Peri. 2006. “The Economic Value of Cultural Diversity: Evidence from US Cities.” *Journal of Economic Geography*, 6(1), 9 - 44.

[72] Pacyga, Dominic A. 2003. *Polish Immigrants and Industrial Chicago: Workers on the South Side, 1880-1922*. Chicago: University of Chicago Press.

[73] Patacchini, Eleonora, and Yves Zenou. 2012. “Ethnic Networks and Employment Outcomes.” *Regional Science and Urban Economics*, 42(6), 938 - 949.

[74] Peoples, James, and Garrick Bailey. 1999. *Humanity: An Introduction to Cultural Anthropology*. Belmont: Wadsworth Publishing.

[75] Petersen, Roger D. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. New York: Cambridge University Press.

[76] Polavieja, Javier G. 2015. “Capturing Culture: A New Method to Estimate Exogenous Cultural Effects Using Migrant Populations.” *American Sociological Review*, 80(1), 166 - 191.

[77] Polland, Annie, and Daniel Soyer. 2013. *Emerging Metropolis: New York Jews in the Age of Immigration, 1840-1920*. New York: New York University Press.

[78] Portes, Alejandro, 1987. “The Social Origins of the Cuban Enclave Economy of Miami.” *Sociological Perspectives*, 30(4), 340 - 372.

[79] 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.

[80] Rubinstein, Yona, and Dror Brenner. 2014. "Pride and Prejudice: Using Ethnic-Sounding Names and Inter-Ethnic Marriages to Identify Labour Market Discrimination." *Review of Economic Studies*, 81(1), 389 - 425.

[81] Russia, Tsentral'nyi statisticheskii komit (Central Statistical Committee of Russia). 1897. *1897 Russian Empire Census*.

[82] Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. *Integrated Public Use Microdata Series: Version 9.0* [Machine-readable database]. Minneapolis: University of Minnesota.

[83] Saiz, Albert. 2007. "Immigration and Housing Rents in American cities." *Journal of Urban Economics*, 61(2), 345 - 371.

[84] Saloutos, Theodore. 1976. "The Immigrant Contribution to American Agriculture." *Agricultural History*, 50(1), 45 - 67.

[85] Sammartino, Annemarie H. 2010. *The Impossible Border: Germany and the East, 1914-1922*. Ithaca: Cornell University Press.

[86] Senik, Claudia, and Thierry Verdier. 2011. "Segregation, Entrepreneurship and Work Values: The Case of France." *Journal of Population Economics*, 24(4), 1207 - 1234.

[87] Simon, Rita J. 1997. *In the Golden Land: A Century of Russian and Soviet Jewish Immigration in America*. Westport: Praeger.

[88] Snyder, Timothy. 2006. "Ukrainains and Poles." In *The Cambridge History of Russia: Volume 2*, ed., Dominic Lieven. New York: Cambridge University Press.

[89] South, Scott J., Kyle Crowder, and Erick Chavez. 2006. "Geographic Mobility and Spatial Assimilation among U.S. Latino Immigrants." *International Migration Review*, 39(3), 577 - 607.

[90] Spielman, Richard S., Laurel A. Bastone, Joshua T. Burdick, Michael Morley, Warren J. Ewens, and Vivian G. Cheung. 2007. "Common Genetic Variants Account for Differences in Gene Expression among Ethnic Groups." *Nature Genetics*, 39, 226 - 231.

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

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

[93] Tabellini, Marco. 2019. "Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration." manuscript.

[94] Thaden, Edward C. 2014. *Russification in the Baltic Provinces and Finland, 1855-1914*. Princeton: Princeton University Press.

[95] Thomas, William, and Florian Znaniecki. 1996. *The Polish Peasant in Europe and America*. Champaign: University of Illinois Press.

[96] Treadgold, Donald. 1957. *Great Siberian Migration*. Princeton: Princeton University Press.

[97] Treeratipuk, Pucktada, and C. Lee Giles. 2005. "Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching." *Twenty-Sixth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5180/5533.

[98] Unbegaun, Boris Ottokar. 1972. *Russian Surnames*. Oxford, U.K.: Oxford University Press.

[99] Waldinger, Maria. 2017. "The Long-Run Effects of Missionary Orders in Mexico." *Journal of Development Economics*, 127, 355 - 378.

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

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

[102] Wetherell, Charles, and Andrejs Plakans. 1999. "Borders, Ethnicity, and Demographic Patterns in the Russian Baltic Provinces in the Late Nineteenth Century." *Continuity and Change*, 14(1), 33 - 56.

[103] Xu, Dafeng. 2017. "Acculturational Homophily." *Economics of Education Review*, 59, 29 - 42.

[104] Xu, Dafeng. 2019. “Surname-Based Ethnicity and Ethnic Segregation in the Early Twentieth Century U.S.” *Regional Science and Urban Economics*, 77, 1 - 19.

[105] Yuengert, Andrew M. 1995. “Testing Hypotheses of Immigrant Self-Employment.” *Journal of Human Resources*, 30(1), 194 - 204.

[106] Zhang, Yan. 2016. “Assessing Fair Lending Risks Using Race/Ethnicity Proxies.” *Management Science*, 64(1), 178 - 197.

[107] Zunz, Olivier. 2000. *The Changing Face of Inequality: Urbanization, Industrial Development, and Immigrants in Detroit, 1880-1920*. Chicago: University of Chicago Press.

Table 1: Demographic Characteristics of Ethnic Populations in Russia

A. 1897 Russian Census by Language and Religion

Language (mother tongue)	Number	Percentage	Religion	Number	Percentage
Russian†	55,667,469	44.31%	Orthodox	87,123,604	69.34%
Pan-Russian languages‡	83,933,567	66.80%			
Yiddish (Jewish language)	5,063,156	4.03%	Jewish	5,215,805	4.15%
German	1,790,489	1.43%	Lutheran⋆	3,572,653	2.84%
Polish	7,931,307	6.31%	Catholics§	11,467,994	9.13%
Turkish-Tatar	13,373,867	10.64%	Muslims	13,906,972	11.07%

B. Major Sending Countries of U.S. Immigrants

Country of birth	1880	1900	1910	1920	1930	1940
Russia∧	32,432	424,647	1,562,045	1,450,734	1,197,244	1,067,956
Italy	44,498	490,883	1,351,055	1,608,841	1,789,588	1,633,659
Germany∧	1,938,065	2,671,484	2,505,650	1,631,480	1,610,701	1,245,601
Ireland	1,853,018	1,641,387	1,355,741	1,049,330	929,429	681,742
England	664,939	850,565	889,485	825,755	813,117	630,001
Canada	716,175	1,229,924	1,254,880	1,2139,53	1,399,034	1,120,505

C. Languages of Russian-Born Immigrants (1930)

Language (mother tongue)	Percentage
Major groups in the U.S.:	
Jewish	62.16%
German	25.04%
Pan-Russian	9.29%
Polish	1.64%
Minor groups in the U.S.:	
Lithuanian	0.57%
Latvian	0.09%
Finnish	0.03%
Turkish-Tatar	0.01%

Notes for Panel A: Source: 1897 Russian census data. †: Not including Ukrainian and Belorussian. ‡: Russian, Ukrainian, and Belorussian. ⋆: 1,435,937 (1.14%) people spoke Latvian. Many Latvians followed Lutheran. §: 1,658,352 (1.33%) people spoke pan-Lithuanian languages. The majority of Lithuanians (as well as some Latvians) followed Roman Catholics.

Notes for Panel B: Sources: 1880-1940 full-count U.S. census data (Ruggles et al., 2019). 1890 census data were destroyed in the 1920s and are thus not available. ∧: Pre-WWI (before 1920) numbers are overestimated due to border changes after WWI.

Notes for Panel C: Sources: 1930 full-count U.S. census data (Ruggles et al., 2019).

Table 2: Ethnic/Religious Segregation in St. Petersburg: Percentages of Non-Orthodox Population That Would Have to Relocate to Achieve Equal Residential Distribution

A. Ethnic/Religious Segregation in St. Petersburg

Religion	1869	1910
Catholics	20.6%	13.5%
Protestants	20.8%	20.0%
Jews	40.7%	52.0%

B. Lawyers' Religions in St. Petersburg Judicial Circuit

	Lawyers		Apprentice lawyers	
Religion	Number	Percentage	Number	Percentage
Orthodox	160	54%	109	41%
Catholics	38	13%	34	13%
Lutheran (Protestants)	36	12%	17	7%
Jewish	62	21%	104	39%
All groups	296	100%	264	100%

C. Ethnic Populations in St. Petersburg's Manufacturing Sector

Position held	Jews	Non-Jews	Jews as % total
Manager ("*khoziaeva*")	29%	9%	5%
Administrator ("*administratsiia*")	4%	3%	N/A
Self-employed ("*odinochka*")	26%	12%	3%
Worker ("*rabochii*")	41%	76%	1%
All positions	100%	100%	

Notes for Panel A: Measure: percentages of non-Orthodox population that would have to relocate to achieve equal residential distribution. Sources: James H. Bater, *St. Petersburg: Industrialization and Change* (1976), and Benjamin *Nathans, Beyond the Pale: The Jewish Encounter with Late Imperial Russia* (2002).

Notes for Panel B: St. Petersburg's Judicial Circuit data in 1888. Sources: Benjamin Nathans, *Beyond the Pale: The Jewish Encounter with Late Imperial Russia* (2002).

Notes for Panel C: St. Petersburg's manufacturing sector data in 1881. Sources: Benjamin Nathans, *Beyond the Pale: The Jewish Encounter with Late Imperial Russia* (2002).

Table 3: Classification Results in Test Data

	Precision	Recall	F-measure	Accuracy
A. Dataset I:				
Russian	0.88	0.86	0.87	
Jewish	0.92	0.90	0.91	
German	0.88	0.98	0.93	
Polish	0.97	0.90	0.93	
Overall				0.91
B. Dataset II:				
Russian	1	0.94	0.97	
Jewish	1	0.98	0.99	
German	0.89	1	0.94	
Polish	0.98	0.94	0.96	
Overall				0.97

Table 4: Classification Results: Percentages of Russian-Born Ethnic Groups in 1930

	(1)	(2)	(3)	(4)	(5)	(6)
Ethnicity:	Russian	Jewish	German	Polish	Unclassified	All
A. Overall %:	0.084	0.622	0.258	0.027	0.009	1.000
B. Year of arrival:						
Before 1900	0.062	0.595	0.313	0.017	0.013	1.000
	[0.130]	[0.170]	[0.215]	[0.115]	[0.255]	[0.177]
1900 - 1904	0.065	0.663	0.245	0.018	0.009	1.000
	[0.125]	[0.175]	[0.156]	[0.183]	[0.164]	[0.164]
1905 - 1909	0.074	0.663	0.235	0.022	0.008	1.000
	[0.200]	[0.244]	[0.208]	[0.186]	[0.198]	[0.229]
1910 - 1914	0.111	0.566	0.271	0.044	0.007	1.000
	[0.353]	[0.245]	[0.283]	[0.444]	[0.224]	[0.269]
1915 - 1919	0.113	0.590	0.250	0.038	0.010	1.000
	[0.054]	[0.039]	[0.039]	[0.057]	[0.045]	[0.041]
1920 - 1924	0.093	0.677	0.207	0.017	0.007	1.000
	[0.108]	[0.108]	[0.079]	[0.063]	[0.078]	[0.099]
1925 - 1930	0.119	0.591	0.246	0.030	0.016	1.000
	[0.030]	[0.020]	[0.021]	[0.023]	[0.037]	[0.022]
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]

Sample size: 609,200 (1930 U.S. full-count census data).

In Panel B, a number in the main line is the fraction of an ethnic group among all Russian-born immigrants who arrived in that specific period (the sum of all numbers in that *row* equals 1).

In Panel B, a number in the bracket is the fraction of people from an ethnic group who arrived in that particular period, among all immigrants from that group (the sum of all numbers in that *column* equals 1).

Table 5: Summary Statistics of Immigrants in 1930 U.S. Census Data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Russian-born immigrants' ethnicity:					Other	Native	(1) - (4)	(5) vs (6)
	Russian	Jewish	German	Polish	All	immigrants	-born	F-test	t-test
A. Demographic variables									
Age	41.609	42.069	43.265	42.244	42.974	42.346	26.792	<0.001	<0.001
	(11.920)	(13.296)	(13.238)	(10.894)	(13.122)	(15.521)	(19.349)		
Years since	21.258	23.096	24.276	21.280	23.197	23.500	—	<0.001	<0.001
migration	(9.632)	(9.627)	(10.877)	(8.775)	(9.984)	(13.516)			
Urban	0.839	0.982	0.718	0.786	0.896	0.780	0.517	<0.001	<0.001
	(0.368)	(0.132)	(0.450)	(0.410)	(0.306)	(0.414)	(0.500)		
Household	3.868	4.412	4.357	3.772	4.334	3.875	4.799	<0.001	<0.001
size	(2.313)	(1.891)	(2.533)	(2.521)	(2.138)	(2.052)	(2.561)		
Citizenship	0.550	0.726	0.641	0.439	0.681	0.569	—	<0.001	<0.001
	(0.490)	(0.446)	(0.480)	(0.496)	(0.466)	(0.495)			
Literate	0.883	0.936	0.913	0.810	0.922	0.903	0.961	<0.001	<0.001
	(0.325)	(0.241)	(0.282)	(0.392)	(0.269)	(0.295)	(0.194)		
Married	0.724	0.798	0.770	0.700	0.781	0.674	0.388	<0.001	<0.001
	(0.447)	(0.402)	(0.421)	(0.458)	(0.413)	(0.469)	(0.487)		
Speak English	0.928	0.954	0.946	0.892	0.948	0.936	0.996	<0.001	<0.001
	(0.259)	(0.209)	(0.227)	(0.310)	(0.222)	(0.245)	(0.063)		
B. Labor market variables									
In labor force	0.909	0.908	0.900	0.906	0.906	0.900	0.872	0.078	<0.001
	(0.288)	(0.289)	(0.301)	(0.291)	(0.292)	(0.300)	(0.334)		
Employed (if in	0.880	0.901	0.899	0.856	0.901	0.875	0.915	<0.001	<0.001
labor force)	(0.325)	(0.299)	(0.301)	(0.351)	(0.299)	(0.330)	(0.279)		
Occupational	26.848	30.631	25.329	24.061	28.694	24.447	22.531	<0.001	<0.001
scores	(10.814)	(10.361)	(11.525)	(8.845)	(10.975)	(9.554)	(11.127)		
Non-agriculture	28.111	30.733	28.952	25.355	29.953	23.354	22.111	<0.001	<0.001
occ. scores	(10.343)	(10.301)	(10.542)	(8.323)	(10.388)	(9.867)	(11.040)		
In higher paying	0.247	0.360	0.230	0.154	0.310	0.243	0.215	<0.001	<0.001
industries†	(0.431)	(0.480)	(0.421)	(0.361)	(0.462)	(0.429)	(0.410)		
Observations	51,418	378,732	157,310	16,319	609,200	7,057,532	54,190,055		

Standard deviations are in parentheses. p-values of statistical tests are shown in Column 8 and Column 9.

Table 6: Summary Statistics: Industries

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Percentage of workers among				Occupational score among		
	all	natives	immigrants	Russians	natives	immigrants	Russians
1. Public administration	2.49%	2.71%	1.44%	0.64%	26.224	24.369	24.432
					(9.630)	(8.294)	(9.474)
2. Agriculture	31.58%	35.07%	15.13%	8.46%	11.699	12.203	12.913
					(3.943)	(3.332)	(3.026)
3. Construction	8.40%	7.82%	11.14%	9.25%	25.991	25.166	26.443
					(6.604)	(6.002)	(6.743)
4. Finance and business	6.07%	6.30%	4.99%	6.11%	29.860	28.383	31.651
					(7.720)	(7.936)	(7.472)
5. Manufacturing	19.07%	17.09%	28.41%	24.23%	25.690	24.957	26.169
					(7.163)	(6.167)	(6.726)
6. Mining	3.37%	3.08%	4.76%	1.46%	25.695	24.573	24.383
					(4.623)	(2.765)	(2.493)
7. Services, professional	3.17%	3.30%	2.54%	3.52%	40.592	35.666	46.796
					(21.480)	(21.360)	(22.798)
8. Services, personal/entertain.	4.13%	3.69%	6.21%	5.57%	20.900	20.176	24.143
					(9.283)	(8.911)	(10.100)
9. Trade, wholesale/retail	12.71%	11.85%	16.74%	36.96%	28.684	29.687	32.102
					(9.046)	(10.487)	(9.596)
10. Transportation/ communication/etc.	9.01%	9.08%	8.65%	3.80%	28.881	25.780	25.097
					(8.283)	(7.375)	(6.539)

Standard deviations are in parentheses. The classification is based on the 1950 industrial classification system (Census Bureau, 1950). Occupational score: median total income (in hundreds of 1950 dollars) of all persons with that occupation in 1950 (Ruggles et al., 2019).

Table 7: Occupational Segregation by Ethnicity among Russian-Born Immigrants

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Percentage of workers among				Occupational scores among			
	Russian	Jewish	German	Polish	Russian	Jewish	German	Polish
1. Public administration	0.88%	0.53%	0.80%	0.88%	20.835	27.635	22.246	19.011
					(10.200)	(8.112)	(9.837)	(7.892)
2. Agriculture	9.23%	0.74%	24.94%	11.26%	12.635	14.290	12.881	12.489
					(3.035)	(6.045)	(2.709)	(3.094)
3. Construction	8.75%	10.56%	7.55%	8.01%	25.381	26.948	25.651	23.992
					(6.501)	(6.819)	(6.527)	(5.571)
4. Finance and business	5.93%	6.76%	5.21%	4.31%	29.454	32.617	30.653	26.816
					(8.376)	(6.880)	(7.893)	(8.466)
5. Manufacturing	28.33%	23.22%	21.30%	38.68%	24.835	27.143	25.450	23.796
					(6.514)	(6.782)	(6.732)	(5.472)
6. Mining	5.36%	0.06%	2.31%	7.95%	24.280	27.821	24.387	24.164
					(2.156)	(7.378)	(2.525)	1.321)
7. Services, professional	3.82%	3.73%	3.26%	2.26%	45.070	47.906	45.297	38.166
					(23.054)	(22.237)	(23.456)	(23.318)
8. Services, personal/entertain.	5.37%	5.83%	4.45%	4.57%	22.773	25.025	23.541	20.790
					(9.792)	(10.283)	(10.046)	(8.880)
9. Trade, wholesale/retail	26.83%	45.88%	24.92%	15.46%	31.846	32.147	32.357	30.111
					(9.866)	(9,503)	(9.680)	(10.066)
10. Transportation/ communication/etc.	5.51%	2.69%	5.20%	6.73%	24.419	25.560	25.073	23.783
					(6.100)	(6.586)	(6.626)	(5.590)

Standard deviations are in parentheses. The classification is based on the 1950 industrial classification system (Census Bureau, 1950). Occupational score: median total income (in hundreds of 1950 dollars) of all persons with that occupation in 1950 (Ruggles et al., 2019).

Table 8: Occupational Segregation by Ethnicity among Russian-Born Immigrants: Regression Analyses

Dep. var.:	Occupational scores				Dummy of higher paying industry†			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Jewish ethnic dummy	3.783** (1.187)	1.156 (0.667)	1.284* (0.638)	0.797** (0.294)	0.198* (0.074)	0.119* (0.057)	0.131* (0.055)	0.012 (0.030)
German ethnic dummy	−1.518 (0.900)	−0.844* (0.346)	−0.124 (0.247)	0.140 (0.110)	−0.031 (0.029)	−0.005 (0.013)	0.025 (0.013)	−0.004 (0.006)
Polish ethnic dummy	−2.787*** (0.339)	−1.657*** (0.176)	−1.929*** (0.149)	−1.065*** (0.141)	−0.145*** (0.007)	−0.112*** (0.009)	−0.128*** (0.007)	−0.043*** (0.004)
Year since migration		0.131*** (0.019)	0.137*** (0.018)	0.112*** (0.014)		0.004*** (0.000)	0.004*** (0.000)	0.004*** (0.000)
Urban		9.970*** (0.841)	7.947*** (0.612)	1.013*** (0.216)		0.313*** (0.017)	0.125*** (0.026)	0.101*** (0.009)
Citizenship		3.179*** (0.151)	3.244*** (0.104)	2.350*** (0.068)		0.098*** (0.006)	0.104*** (0.006)	0.076*** (0.005)
Other controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
State FE	No	No	Yes	Yes	No	No	Yes	Yes
Industry control	No	No	No	FE	—	—	w/o ag.	w/o trade
Adjusted R^2	0.052	0.177	0.199	0.388	0.053	0.110	0.076	0.061
Observations	441,278	441,278	441,278	378,025	396,233	396,233	362,936	247,229

(1) - (4): Regressions of earnings. Reference group: Russian ethnics (the majority group in Russia).
(5) - (8): Regressions of the propensity to work in higher paying industries (†: finance/business, professional services, and trade) by ethnicity.
Standard errors are in parentheses and are clustered at the state level. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Table 9: Ethnic Occupational Segregation in the U.S. and Russia

A. Manufacturing Jobs by Ethnicity in St. Petersburg and the U.S.

Position held, St. Petersburg	Jews	Non-Jews	Ratio	Position held, U.S.	Jews	Non-Jews	Ratio
Manager	29%	9%	3.2:1				
Administrator	4%	3%	1.3:1				
Manager + Administrator	33%	12%	2.7:1	Employer	7.03%	3.06%	2.3:1
Self-employed	26%	12%	2.2:1	Self-employed	2.25%	4.41%	2:1
Manager + Admin. + Self-emp.	59%	24%	2.4:1	Employer + Self-emp.	11.44%	5.31%	2.2:1
Worker	41%	76%	1:1.9	Works for wage	88.56&	94.85%	1:1.1

B. Jews' Occupations in Russia and the U.S.

Industry among workers	Russia	U.S.
Agriculture	3.76%	0.74%
Administration	5.53%	0.53%
Manufacturing	37.49%	23.22%
Service	6.99%	5.83%
Trade & Commerce	40.90%	45.88%
Transportation	4.21%	2.69%

Notes for Panel A: Data sources: James H. Bater, *St. Petersburg: Industrialization and Change* (1976) and 1930 U.S. census data (Ruggles et al., 2019).

Notes for Panel B: Data sources: Rita J. Simon, *In the Golden Land: A Century of Russian and Soviet Jewish Immigration in America* (1997) and 1930 U.S. census data (Ruggles et al., 2019).

Table 10: The Relationship between the Spatial Concentration of Ethnicity and Labor Market Outcomes

Dependent var.:	# same (1), diff. (2) ethnics		Employment status		Occupational scores		Higher paying industry	
	First-stage regressions		OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Historical settlements of immigrants (IVs):								
# (in k) in the county,	1.358***	1.969***						
same ethnicity	(0.334)	(0.365)						
# (in k) in the county,	0.441**	3.250***						
different ethnicity	(0.166)	(0.193)						
Enclave of old immigrants (arrived before 1920):								
# (in k) in the county,			0.0018***	0.0091*	0.0763*	0.3498*	0.0084***	0.0299*
same ethnicity			(0.0003)	(0.0045)	(0.0330)	(0.1508)	(0.0004)	(0.0075)
# (in k) in the county,			0.0015***	−0.0003	0.0778*	0.0416	0.0005	−0.0040
different ethnicity			(0.0002)	(0.0012)	(0.0356)	(0.0421)	(0.0004)	(0.0021)
Enclave of new immigrants (arrived in/after 1920):								
# (in k) in the county,			−0.0119**	−0.0553*	−0.5376**	−2.1685*	−0.0583***	−0.1872***
same ethnicity			(0.0016)	(0.0027)	(0.1951)	(0.9035)	(0.0025)	(0.0451)
# (in k) in the county,			−0.0095*	0.0011	−0.5495**	−0.3057	−0.0127***	−0.0150
different ethnicity			(0.0014)	(0.0076)	(0.1792)	(0.2585)	(0.0022)	(0.0129)
County population	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F	> 100	> 100						
Adjusted R^2	0.327	0.260	0.069	—	0.164	—	0.107	—
Observations	492,149	492,149	609,200	492,149	445,099	356,127	399,661	318,050

Dependent variables: (1): the number of immigrants of same ethnicity in the local county; (2) the number of immigrants of different ethnicity in the local county; (3) - (4): employment status; (5) - (6): occupation-based earnings; (7) - (8): the indicator of working in a higher-paying industry.
Independent variables: numbers of immigrants (in k) in the local county, by ethnicity (same or different).
Instrumental variable: historical settlements of immigrants by ethnicity (in 1880).
Higher paying industries: finance/business, professional services, and trade.
Standard errors are in parentheses, and are clustered at the state level. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Table 11: The Relationship between the Spatial Concentration of Ethnicity and Labor Market Outcomes: Compatriot (Co-Birth Immigrants) Networks

Dependent var.:	Emp. status	Occ. scores	Higher paying ind.
	IV	IV	IV
	(1)	(2)	(3)
Enclave of old immigrants:			
# (in k) in the county,	0.0017***	0.1071***	0.0038**
born in Russia	(0.0004)	(0.0249)	(0.0012)
Enclave of new immigrants:			
# (in k) in the county,	−0.0105**	−0.7130***	−0.0310**
born in Russia	(0.0025)	(0.1510)	(0.0083)
County population	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	492,149	356,127	318,050

Independent, dependent, and instrumental variables follow the previous table.

Higher paying industries: finance/business, professional services, and trade.

Standard errors are in parentheses and are clustered at the state level. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Table 12: The Relationship between the Spatial Concentration of Ethnicity and Labor Market Outcomes: Heterogeneous Effects

Dep. var.:	Employment status				Occupational scores				Higher paying industry			
Ethnic group:	RUS	JEW	GER	POL	RUS	JEW	GER	POL	RUS	JEW	GER	POL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Enclave of old immigrants:												
# (in k) in the county,	0.0429***	0.0040***	0.0161***	0.0049	0.4789	0.2804***	2.7617***	0.0136***	0.0008**	0.0305***	0.0832***	0.0008***
same ethnics	(0.0109)	(0.0008)	(0.0025)	(0.0045)	(0.3245)	(0.0337)	(0.0745)	(0.0037)	(0.0002)	(0.0060)	(0.0034)	(0.0001)
# (in k) in the county,	0.0013	−0.0052	0.0060***	0.0043*	0.0020	0.0558	0.2167***	0.0084**	−0.0003**	−0.0122***	0.0145***	−0.0005*
different ethnics	(0.010)	(0.0031)	(0.0008)	(0.0020)	(0.0298)	(0.0302)	(0.0239)	(0.0026)	(0.0001)	(0.0028)	(0.0012)	(0.0002)
Enclave of new immigrants:												
# (in k) in the county,	−.0727***	−0.0226***	0.0842***	0.0528	0.4515	−1.5192***	−8.9961***	−1.2373***	−2.9812**	−0.1459***	−0.3390***	−0.2623***
same ethnics	(0.0205)	(0.0049)	(0.0091)	(0.0312)	(0.6064)	(0.2121)	(0.2771)	(0.3545)	(0.9570)	(0.0041)	(0.0128)	(0.0372)
# (in k) in the county,	−0.0143*	0.0016	−0.0334***	−0.0250*	0.0939	−1.4469***	−0.5421***	−0.5205**	1.4656**	−0.0371***	−0.0691***	−0.0003*
different ethnics	(0.0061)	(0.0077)	(0.0049)	(0.0121)	0.1815	(0.3357)	(0.1489)	(0.1644)	(0.4637)	(0.0072)	(0.0069)	(0.0001)
County population	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	51,418	378,732	157,310	16,319	38,441	272,002	118,372	12,463	28,223	202,180	108,263	11,417

Independent variables: numbers of immigrants (in k) in the local county. IV: historical settlements of immigrants by ethnicity (in 1880).
Higher paying industries: finance/business, professional services, and trade. Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

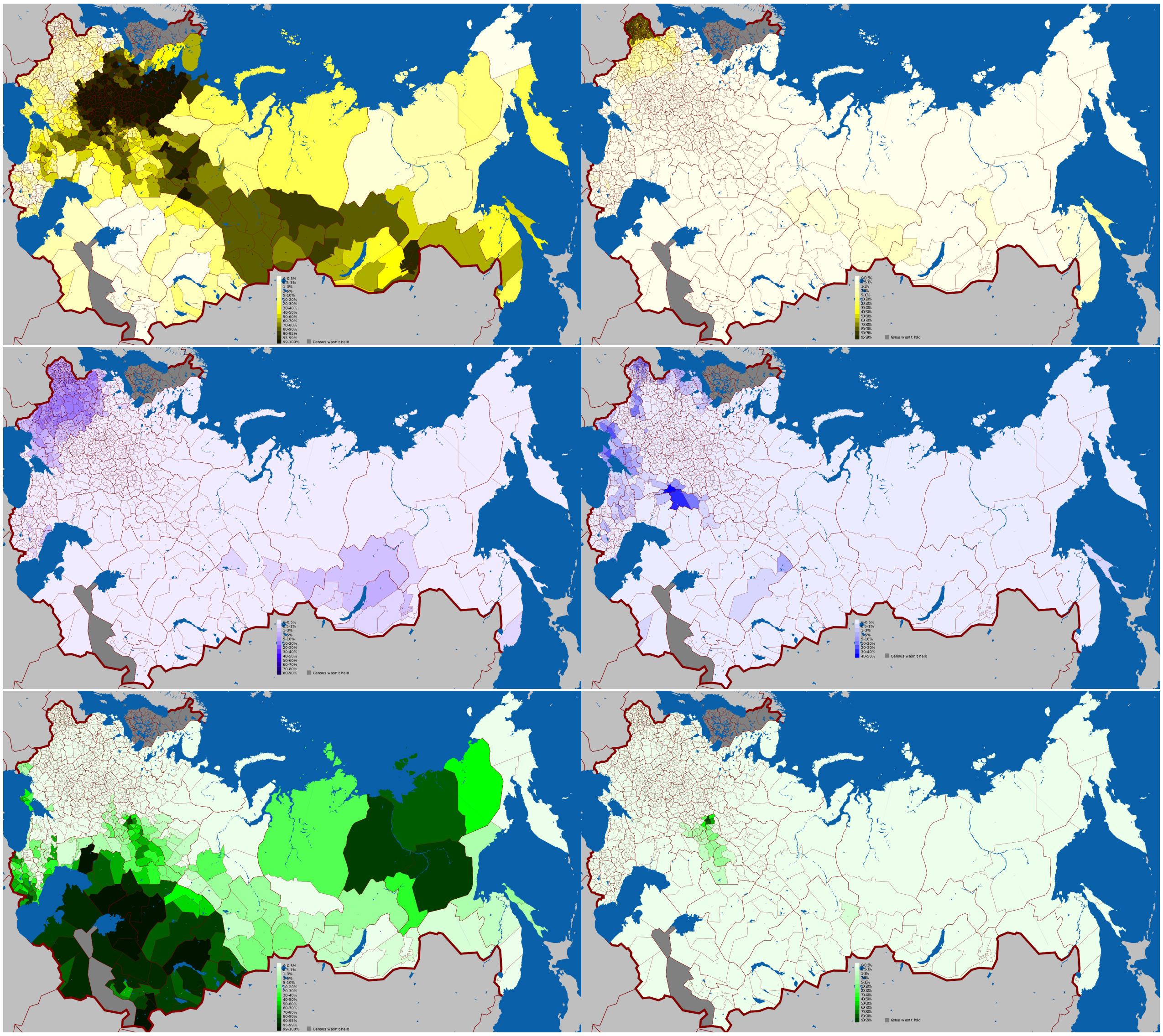

Figure 1: Geographic Distribution of Language Groups in Imperial Russia, 1897 (Source: 1897 Russian Census Data)
Row 1: Slavic Languages. Left: Russian. Right: Polish.
Row 2: Germanic Languages. Left: Yiddish. Right: German.
Row 3: Turkic Languages. Left: Turkic Languages (All Types). Right: Chuvash.

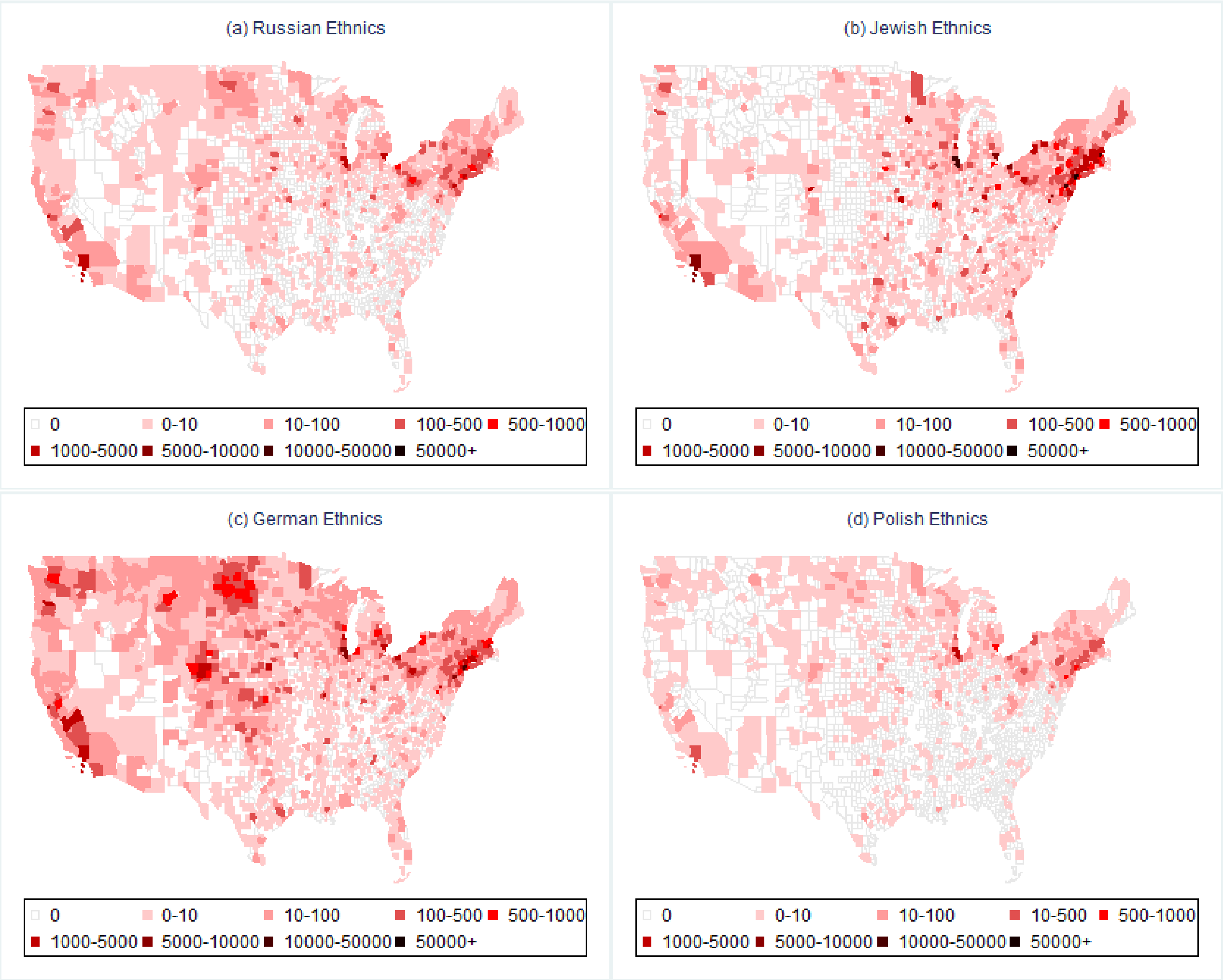

Figure 2: Geographic Distribution of Ethnic Groups among Russian-Born Immigrants in the U.S.