The Structure and Growth of Ethnic Neighborhoods*

Tianran Dai† Nathan Schiff‡

January 2023

Abstract

We introduce a new statistical definition of an immigrant ethnic neighborhood based on a choice model and using the location distribution of natives as a benchmark. We then examine the characteristics of ethnic neighborhoods in the United States using decadal census tract data from 1970-2010. We estimate that 43% of the foreign-born population lived in ethnic neighborhoods in 1970, increasing to 67% by 2010. Ethnic neighborhoods have lower average incomes and housing values, and a higher percentage of residents living in rental housing and commuting without a car, than other locations in the city where the same group lives. Neighborhoods vary greatly in size and the population distribution across neighborhoods within a group follows a power law. Most neighborhoods disappear within one or two decades but larger neighborhoods persist longer. Large neighborhoods have a well-defined spatial structure with negative population gradients and growth occurs primarily through spatial expansion into adjacent locations.

JEL codes: R23, R30, J15

Keywords: neighborhoods; ethnic enclave; concentration metrics; housing

Maps of ethnic neighborhoods: https://nathanschiff.shinyapps.io/ee_maps_deployed/

1 Introduction

In 2010, over half of New York City's 425 thousand China-born residents lived in just 207 of the city's 4661 census tracts, less than 4.5% of the total. Additionally, a single cluster of 26 contiguous tracts accounted for 9.2% of the entire city's Chinese population. As we describe in detail later, this level of geographic concentration is far greater than that of the native-born population, or the general population across all groups, in 2010 New York City. Yet this phenomenon is by no means rare. Across the United States and over the five decades of 1970-2010 we study, we find similarly high concentration levels for many different immigrant groups in many cities. Moreover, these numbers would likely surprise few city residents; ethnic neighborhoods, such as "Little Italy" or "Chinatown," have been a common feature of US cities for well over a century. In this paper we provide a statistical definition of an ethnic neighborhood and describe characteristics of these neighborhoods across ethnic groups defined by country of birth. We then study the dynamics of these neighborhoods: what is their spatial structure, how long do they last, and how do they grow?

*We are grateful to the Canadian Social Sciences and Humanities Research Council, SSHRC grant #430-2013-0812, for generously funding our research. We also especially thank Sanghoon Lee for considerable help and advice throughout this project.

†School of Economics, Shanghai University of Finance and Economics

‡School of Economics, Shanghai University of Finance and Economics. Corresponding author: nschiff@gmail.com.

Recently, there has been a resurgence of interest in the causal effects of neighborhoods on the outcomes of residents (Chetty, Hendren, and Katz 2016, Chetty and Hendren 2018a, Chetty and Hendren 2018b). These papers emphasize the importance of place effects, arguing that the neighborhood in which a child grows up has a causal effect on lifetime earnings and other important outcomes. The findings closely relate to earlier analyses of ethnic neighborhoods, which show evidence for place effects that also vary by ethnicity. This literature finds important effects on wages, educational attainment, language acquisition, and the likelihood of finding employment through ethnic job networks.[1] A separate literature in public health examines the effect of ethnic neighborhoods on health and medical outcomes, including diet, smoking, diabetes, hypertension, and breast cancer (Osypuk, Roux, Hadley, and Kandula 2009, Keegan, John, Fish, Alfaro-Velcamp, Clarke, and Gomez 2010, Lim, Stella, De La Cruz, and Trinh-Shevrin 2017). Further, while place effects can be difficult to identify when location choices are endogenous, a number of papers have used refugee placement programs as a source of exogenous variation in location and found persuasive evidence that ethnic concentration can affect education and labor market outcomes.[2] Thus a fairly extensive body of work suggests that neighborhoods—and perhaps especially ethnic neighborhoods—can have important causal effects on both children and working age adults.

In contrast to the existing literature on the outcomes of neighborhood residents, this paper focuses on the neighborhoods themselves, using census tracts as a unit of analysis, rather than individuals. We make three contributions. First, we provide a statistical definition of an ethnic neighborhood derived from a choice model of location and show that our method better predicts outcomes commonly associated with ethnic neighborhoods than simpler definitions. Second, we provide the first comprehensive description of ethnic neighborhoods in the United States over five decades and use this to examine fundamental questions from the U.S. immigration literature. Third, we examine the structure and growth of ethnic neighborhoods—an area with little to no previous work—and find that ethnic neighborhoods have striking similarities to empirical patterns documented in and across cities.

We base our definition of an ethnic neighborhood on a discrete choice model where members of an ethnic group choose residences among a city's census tracts. Using the total population as a proxy for the housing capacity of any location, we show that the share of the total ethnic population choosing a location divided by the share of the total population choosing that location can be interpreted as a measure of the ethnic-group specific utility of a location, relative to all others in the city. This ratio of shares is often used to measure firm or industry concentration and is known as the location quotient (LQ). We use natives—US-born residents—as a baseline for comparison, and then define an ethnic (group) tract as any census tract where the location quotient of that ethnic group is statistically greater than the 99th percentile of the native location quotient distribution in that city-year. The intuition for this measure is that it allows us to identify concentrations of an immigrant group that would be extremely unusual if members of an ethnic group had the same distribution of relative utility over locations as natives, but without the assumption that the relative utility of a specific location is the same for natives and ethnic group members. Having identified ethnic tracts, we then group together clusters of contiguous tracts and label these as ethnic neighborhoods.

An important body of work in sociology has documented the characteristics of ethnic neighborhoods,

[1] Borjas (1995) finds that the wages and educational attainment of children may be affected by the human capital level of co-ethnic adults in the same neighborhood. Cutler, Glaeser, and Vigdor (2008b) find that larger ethnic populations increase the earnings of young workers and the English language ability of young adults of the same ethnicity in a census tract. Hellerstein, Kutzbach, and Neumark (2014) provide evidence for spatial job networks at the census tract level operating through ethnicity (see also (Munshi 2003) and (Ioannides and Loury 2004) on ethnic job networks).

[2] See the article by Edin, Fredriksson, and Åslund (2003), as well as Damm (2009, 2014),and Aslund, Edin, Fredriksson, and Gronqvist (2011). A recent paper by Abramitzky, Boustan, and Connor (2020) studies the effect of leaving an enclave by examining the case of a (non-refugee) program in the early 20th century that helped Jewish households living in New York City move to non-enclave locations.

using context-specific knowledge of locations or a variety of statistical methods to define neighborhoods (see the recent survey by Gold (2015) or chapter 3 of Portes and Rumbaut (2014)). However, much of this evidence is based on case studies of specific groups, either cross-sectional or confined to select years, and using neighborhood definitions that are difficult to compare across studies. Given the heterogeneity in immigrant location patterns, this approach is helpful to understand specific groups, but it cannot provide answers to more general questions about how immigrant neighborhoods have changed over time. What percentage of the foreign-born live in ethnic neighborhoods in the United States, and is this percentage increasing or decreasing? Has the average ethnic neighborhood significantly decentralized and moved out of the city to the suburbs, consistent with several well-studied examples of "ethnoburbs" (Li 1998, Logan, Zhang, and Alba 2002, Zhou, Tseng, and Kim 2008), or are these examples exceptional? How long does the average ethnic neighborhood last?

In this paper we calculate a set of statistics for the average (ethnic) resident of an ethnic neighborhood—weighted across all foreign-born groups—and for every census decade, 1970-2010. We estimate that 43% of the foreign-born population lived in ethnic neighborhoods in 1970, with this percentage increasing every decade to 67.2% in 2010. The ethnic population in an ethnic tract has also increased considerably from an average of 384 people in 1970 to 683 people in 2010. The median population of an ethnic neighborhood—a cluster of ethnic tracts—has increased from 1,858 (co-ethnic) people in 1970 to 4,032 people in 2010. Consistent with the case studies in the literature, we find that ethnic neighborhoods have decentralized considerably, with a 56% increase in the average distance to the city center between 1970 and 2010, and a 40% decrease in the population density (measured using the total population, ethnic and non-ethnic). The average census tract in an ethnic neighborhood, compared to other tracts in the same city and year where members of the ethnic group live, has significantly lower median household income, a higher percentage of rental housing, an older housing stock, and a larger share of residents commuting without a car. Perhaps reflecting decentralization, the percentage of rental housing in the average ethnic neighborhood and the share of residents commuting without a car have both declined significantly and monotonically since 1970. However, there has been no corresponding change in household income: the household income in an ethnic neighborhood was about $9,000 (2010 dollars) less than in other locations housing the same group, in all five decades. On duration, we find that nearly two thirds of ethnic tracts disappear after a decade, however much of this reflects small, isolated tracts and bigger neighborhoods last longer. Calculating the transition probability for a given resident of an ethnic tract, thus putting more weight on tracts with more people, the likelihood that a tract is still an ethnic tract a decade later is about 70% and drops by roughly ten percentage points in each subsequent decade. The number of large ethnic neighborhoods (a co-ethnic population of at least 1000) for some European groups, including Germans, Hungarians, Irish, and Italians, has decreased significantly since 1970, while the number of neighborhoods for some Asian countries (China, India, Korea, Vietnam) has increased.

We present several new findings on the structure and growth of ethnic neighborhoods. First, ethnic neighborhoods have a spatial structure similar to city sub-centers, with a population that declines from a central point and geographic growth through expansion into neighboring tracts. As a result, the year in which a tract became part of a large neighborhood increases with distance to the neighborhood center. Second, also echoing the empirical literature on cities, the size distribution of ethnic neighborhoods for a given group follows a power law. Third, after controlling for the ethnic population, ethnic neighborhood status and the population of the neighborhood are significant predictors of future ethnic population growth in the neighborhood, and in locations adjacent to the neighborhood. Lastly, we find that ethnic neighborhoods grow in population primarily through spatial expansion, rather than further inflows into existing locations.

Our work is related to a literature on measuring the spatial extent of economic activity, especially defining the spatial boundaries of cities based on economic measures rather than political jurisdictions.[3] As those papers argue, more accurate measurement of spatial boundaries is helpful for any analysis across spatial units (ex: cities, markets, neighborhoods). Papers on ethnic concentration often use a spatial definition of an ethnic neighborhood as a descriptive tool to represent cases with high ethnic concentration. This discrete classification is useful because it simplifies the bivariate distribution between ethnic concentration and other variables of interest, such as education or income, allowing authors to easily compare the characteristics of locations with high versus low ethnic concentration. For example, Abramitzky, Boustan, and Connor (2020) define Jewish enclaves in 1910 New York City as enumeration districts that were at least 40% Jewish and use residents of these enclaves as a comparison group for a Jewish resettlement program. Edin, Fredriksson, and Åslund (2003) define enclaves as municipalities where the ratio of an ethnic population to the general population was at least twice as large as the corresponding ratio in the country (Sweden), and then use this definition to show that individuals placed in enclave municipalities by a refugee resettlement program were comparable to those placed in other municipalities. Several papers in the public health literature define enclaves using various indices or quantiles of population shares.[4] In a similar context to ethnic neighborhoods, Cutler and Glaeser (1997) define Black American ghettos as cities with a dissimilarity index greater than 0.6 and an isolation index greater than 0.3, but without specifying the exact tracts that comprise the ghetto. An implicit assumption behind many of these studies is that important outcomes only occur, or are only observable, when the population concentration of a group passes a given threshold. Therefore, one aim of this paper is to provide a more systematic and micro-founded definition of a threshold that other researchers can use to study ethnic concentration effects.

Additionally, defining ethnic neighborhoods can help to capture important spatial effects. Outcomes in a geographic unit may be affected by populations in nearby units, and this effect may not be a linear function of distance; ethnic neighborhoods may have irregular shapes not captured by simple adjacency measures. In fact, we find that controlling for the ethnic population of a tract, our measure of neighborhood status still has a significant effect on predicting ethnic population growth, birth-country language use, and English ability. Moreover, controlling for both the ethnic population of a tract and that of immediately adjacent tracts, the size of an ethnic neighborhood (across multiple contiguous tracts) is still a significant predictor of the focal tract outcomes.

In the next section we define ethnic neighborhoods using a household-level choice model of residence. In the following section we describe the size, prevalence, and distribution of these neighborhoods, along with how their demographic characteristics have changed over time. We then explore the spatial structure of ethnic neighborhoods in Section 4 and neighborhood dynamics—persistence and growth—in Section 5. We conclude with brief thoughts for future research.

[3] This is a large literature. For very recent work see De Bellefon, Combes, Duranton, Gobillon, and Gorin (2020), Vogel, Goldblatt, Hanson, and Khandelwal (2020), Moreno-Monroy, Schiavina, and Veneri (2020), Dingel, Miscio, and Davis (2021), and other articles in the same special issue of the *Journal of Urban Economics*.

[4] Osypuk, Roux, Hadley, and Kandula (2009) define immigrant enclaves as neighborhoods where the percentage of foreign-born is in the highest quartile and then investigate the effect of living in a Hispanic or Chinese neighborhood on diet and exercise. Keegan, John, Fish, Alfaro-Velcamp, Clarke, and Gomez (2010) uses a collection of mostly language variables to define "high" and "low" enclave neighborhoods and studies breast cancer incidence among Hispanic women. Lim, Stella, De La Cruz, and Trinh-Shevrin (2017) defines enclaves using functions of segregation indexes (dissimilarity, isolation) to study the association between enclave residence and smoking, hypertension, diabetes, and health perception.

2 Defining Neighborhoods

There is no universally agreed upon definition of an ethnic neighborhood. The sociology literature generally defines ethnic enclaves as ethnic population concentrations in employment or residence.[5] The "spatial assimilation" theory suggests that new immigrants with few resources, difficulties speaking the local language, and/or little knowledge of a new country concentrate in neighborhoods with co-ethnics as a type of mutual support system. The ethnic neighborhood is typically located in the central part of the city to minimize commuting costs, provides access to jobs and housing through ethnic networks, supports group-specific cultural institutions, has stores and shopping catering to the group, and operates in the origin-country language. As immigrants gain language abilities, cultural skills, and resources, they move out of the ethnic neighborhood and their spatial patterns approximate those of natives (Allen and Turner 1996, Alba, Logan, Stults, Marzan, and Zhang 1999, Logan, Zhang, and Alba 2002). A seminal paper by Logan, Zhang, and Alba (2002) notes that while this process describes many groups immigrating to the US in the early 20th century, some more recent immigrants with professional skills and substantial resources have clustered in suburban locations with relatively high housing values. Rather than being constrained by limited financial resources and English ability, these immigrants may have chosen concentrated areas in the suburbs for ethnic-specific amenities. They refer to these neighborhoods as "ethnic communities" and define them using a spatial correlation technique with census tract data in New York and Los Angeles.[6]

Similar to Logan, Zhang, and Alba (2002), we set aside the different theories for neighborhood formation and instead define an ethnic neighborhood statistically as a geographic location where there is an unusual residential concentration of an ethnic population. Most of the literature on the effects of ethnic concentration studies outcomes that occur at high levels of residential ethnic population concentration. In addition to the studies mentioned earlier, there is also research on many other outcomes including welfare participation (Bertrand, Luttmer, and Mullainathan 2000), language use (Lazear 1999), housing prices (Saiz and Wachter 2011, Wong 2013), real estate transactions (Agarwal, Choi, He, and Sing 2019), rental discrimination (Bosch, Carnero, and Farre 2010), and the presence of ethnic restaurants (Schiff 2015). We expect that outcomes affected by ethnic concentration will be much more likely to occur in unusually concentrated areas and therefore focus on identifying such locations.

The next question is how to define an unusual concentration. Should concentration be defined as the fraction within a geographic unit (ex: ethnic group g comprises 10% of the total population of tract j) or the fraction of the total population located in that unit (ex: 10% of the city-level population of ethnic group g lives in tract j)? Should the concentration level be deflated by some general population measure in order to compensate for differences in the size of the total ethnic populations?[7] Does it matter that most individuals

[5] Portes and Jensen (1989) argues for a place of work definition while Sanders and Nee (1992) argue for a residence definition—also see Waldinger (1993) for discussion of the debate.

[6] Our paper also uses census tract measures of the ethnic population to define neighborhoods, but differs significantly in theoretical motivation, method, and inference. Logan, Zhang, and Alba (2002) define ethnic tracts as those where the tract and neighboring tracts have above average ethnic population percentages, which is the ethnic population of the tract divided by the total population of the tract. They implement this idea by calculating a measure of local spatial correlation known as the local Moran's I (see Anselin (1995)), and use the associated methods for inferring statistical significance. This method is adept at finding unusual clusters of contiguous tracts, but in this paper we are interested in studying neighborhood growth and thus require a method that can also identify a singleton tract as unusual. Relatedly, our paper uses data approximated to constant boundary census tracts, but concentration measures based on contiguity are not robust to changes in spatial boundaries. A single tract with a high ethnic population may be uncorrelated with its neighbors, but breaking it into four component pieces will create local spatial correlation. Lastly, the tract percentage is a function of the total population, and thus tracts with larger populations (perhaps due to greater housing stock) are less likely to be classified as ethnic tracts.

[7] As an example, Bertrand, Luttmer, and Mullainathan (2000) define a network measure as the proportion of the neighborhood speaking a language, divided by the proportion in the country speaking the language, to prevent "underweighting small language

live in households with other people? These types of issues have been considered at length in the literature on constructing indices for the geographic concentration of manufacturing and firms.[8] Therefore we follow this literature and derive our definition of an unusually concentrated location from a discrete choice model of residence.[9]

2.1 Choice Model of Residence

Since most individuals live in multi-person households, we model the choice of a household over all the H houses in a city, grouped into J locations. Household i in group g receives the following indirect utility from choosing house h in location j:

$$V_{igjh} = ln(\theta_{gj}) + \varepsilon_{igjh} \tag{1}$$

The $ln(\theta_{gj})$ is the utility of any house in location j to any group g household, and represents group-level preferences and constraints, including financial constraints or housing discrimination. We assume that the number of houses in j, H_j, is independent of the choices of any group, and thus we interpret housing as exogenous and an approximation of a location's residential capacity. The ε_{igjh} term is an idiosyncratic match between house h and household i, which we assume is distributed type 1 extreme value across all houses in the city.[10] We do not observe location choices over houses, and thus we assume all housing units in location j are identical, except for the match value ε_{igjh}. If there are J total locations in the city, then the probability a household in group g chooses a *specific* housing unit h in location j is:

$$P_{gjh} = \frac{exp(ln(\theta_{gj}))}{\sum_{k=1}^{J}\sum_{h=1}^{H_k} exp(ln(\theta_{gk}))} = \frac{\theta_{gj}}{\sum_{k=1}^{J} H_k\theta_{gk}} \tag{2}$$

The probability that a household in group g chooses to live in *any* of the H_j houses in location j is therefore just the sum of the (identical) probabilities of living in each house: $P_{gj} = H_j * P_{gjh}$. We can estimate the probability P_{gj} with the share of group g observed to have chosen location j, $s_{gj} = n_{gj}/N_g$, where n_{gj} is the count of group g households in location j and N_g is the total group g household count in the city. For two groups A and B, equation 2 implies that if the share of group A choosing location j is greater than the share of group B choosing a different location k, $s_{Aj} > s_{Bk}$, then group A has stronger relative utility for location j than B has for k, or there are more houses in j than k, or some combination of both. If there are $H = \sum_j H_j$ total houses in the city and a household is indifferent between all houses, then the probability of choosing any particular house is just $1/H$. If we divide P_{gjh} by $1/H$ then we have a ratio comparing the probability a household in group g chooses a particular housing unit to a uniform distribution over all houses:

$$\frac{P_{gjh}}{1/H} = \frac{P_{gj}}{H_j/H} \tag{3}$$

We have assumed identical houses within a location, but of course, the actual number of people who can live in a house varies across units. To avoid having to assess the housing capacity of a location, we instead

groups."

[8]Influential papers include Ellison and Glaeser (1997), Mori, Nishikimi, and Smith (2005), Duranton and Overman (2005), and Guimarães, Figueiredo, and Woodward (2007).

[9]Many papers model residential location using discrete choice models; see Bayer, Ferreira, and McMillan (2007) for a prominent example. Similar to our paper, Billings and Johnson (2012b) define a measure of unusual industrial concentration using a choice model. We discuss this paper at length below.

[10]If we allowed for correlation across houses within the same location then this would generate nested logit choice probabilities, which are too complicated for our application.

proxy for the housing stock in a location with the total population. Glaeser and Gyourko (2005) document that there is nearly a perfect correlation between housing units and population, and in fact theoretical models of housing supply often assume that the number of housing units equals the population, such as the model in Saiz (2010). Denote the total population across all groups in j as n_j, the total population in the city as N, and the share in location j as $s_j = n_j/N$. Substituting the group shares for the group probabilities and the total population shares for the housing shares in equation 3 yields:

$$LQ_{gj} = \frac{s_{gj}}{s_j} = \frac{P_{gj}}{n_j/N} = N * P_{gjh} = \frac{\theta_{gj}}{\sum_{k=1}^{J} \frac{n_k}{N}\theta_{gk}} = \frac{\theta_{gj}}{\sum_{k=1}^{J} s_k\theta_{gk}} \tag{4}$$

In urban economics, the share of a group choosing a location divided by the total population share choosing the location is known as the "location quotient," or LQ. This measure is usually used in studies of industrial location[11] where each group is an industry and the agents making decisions are individual firms or employees (Billings and Johnson 2012a, Guimarães, Figueiredo, and Woodward 2009). The share of all firms choosing a location is used as a proxy for economic activity, under the assumption that in the absence of specific industry-location benefits, firms should generally locate in proportion to economic activity. If the location quotient for an industry is greater than one, it indicates that the concentration of firms from that industry is high relative to the economic activity of the location. In our model of residential location choice, and given the earlier assumptions, the location quotient has a precise meaning defined in terms of group-level utility. It is equal to the (exponentiated) group-level utility of a house in location j, θ_{gj}, divided by the average group-level utility across all houses in the city. The advantage of the location quotient over the simpler population share is that it provides a measure of relative utility that adjusts for the capacity constraints of all locations. Thus, if the LQ for group A is larger in location j than for group B in location k, it implies that the relative utility of A living in a house in j is greater than the relative utility of B living in a house in k.[12]

2.2 Choosing a Threshold

Having defined a measure of concentration, the next step is to choose a threshold concentration level above which a location is classified as an ethnic group location. To do so requires comparing the group g distribution to a reference distribution and then testing whether the concentration level in a given location is statistically larger than would be expected under the null hypothesis that the distributions are equal. The choice of reference distribution is quite important because the interpretation of a group g location stems from the contrast with the reference group: in these locations the concentration levels of group g exceed that of the reference group due to factors that have greater influence on group g. In the firm location literature, several influential papers use the overall distribution of firms, or the distribution of firms within a particular sector, as the reference distribution for firms in a more granular category (Ellison and Glaeser 1997, Duranton and Overman 2005, Billings and Johnson 2012b). For example, Billings and Johnson (2012b) define a location j as "specialized" for an industry g if the density of industry g firms in j is statistically larger than the density of all firms in the sector, in j. In our context we could define "specialized locations" by testing the null hypothesis that the expected group share in a location is equal to the total population share, implying

[11] A recent paper by Louf and Barthelemy (2016) derives this same measure by showing that if a population is disaggregated into groups and allocated to a set of areal units with a fixed total population, then the group population in a specific unit must follow a binomial distribution. They use this measure to explore segregation by income in the United States.

[12] The choice model also illustrates the issue with using the ethnic fraction within a geographic unit: n_{gj}/n_j. This is equivalent to $N_g * P_{gj}/n_j = N_g * P_{gjh} * n_j/n_j = N_g * P_{gjh}$. Thus independent of preferences (P_{gjh}), the fraction within a geographic unit is increasing in the city level population of group g, N_g.

the location quotient is equal to 1:

$$E[s_{gj}] = s_j \iff E[LQ_{gj}] = 1 \tag{5}$$

This approach has been successful in studies of industrial concentration, but the location distribution of the total population may be less relevant for the patterns of specific ethnic groups than the overall concentration of firms is for specific industries. In panel A of Figure 2 we plot a subset of the New York City Core Based Statistical Area (CBSA)—a U.S. Census geography consisting of an urban area and economically integrated adjacent counties—containing 72% of the census tracts and 89% of the 2010 Chinese population. Each dot represents a census tract centroid—the longitude and latitude of the geographic center—and we color the dots according to the count of China-born residents, ranging from 0 (light gray) to 7649 (bright red). Since census tracts have roughly equal population, the dot density is very high in Manhattan and the neighboring boroughs and then decreases considerably as tract population density declines and geographic area increases. Most New York City tracts have very few Chinese residents: 1,947 of the 4,661 tracts do not have a single resident born in China. Some of these zeros may be the result of measurement error in the census data (discussed later in section 3.1), but even very low counts in such a large number of locations is extremely unlikely if Chinese residents locate in a similar way to that of the entire population.[13] On the other hand, there are several clusters with counts that are orders of magnitude larger than most other locations. Again, if Chinese patterns follow the overall population distribution, then the likelihood of observing tracts with such a high concentration is also essentially zero. While there is a statistically significant association between Chinese share and total population share in New York City, a regression of Chinese share on total population share yields an R-squared of less than 0.03 (coefficient is 1.16 with standard error of 0.10). These types of differences in the location patterns of ethnic groups and the total population are found across many groups, cities, and years.

Moreover, using the overall population distribution as a threshold—an LQ of 1—fails to capture the important differences between immigrant concentration and native concentration. In fact, defining natives (individuals born in the United States) as an ethnic group, we find that nearly 50% of the census tracts in every city have a location quotient statistically greater than one. As noted earlier, the general notion of ethnic neighborhoods is that the unusual levels of concentration result from factors specific to immigrant groups, such as speaking a different language, using ethnicity-based job or housing networks, participating in particular community institutions, or purchasing goods with low demand from consumers outside the group. Therefore, to try and capture immigrant specific reasons for concentration we use natives as the reference group, and thus concentration levels of an ethnic group above those of natives may reflect factors specific to immigrants.[14] However, we do not want to compare the preferences and constraints of natives and group g for the same location; these are very unlikely to be equal, and perhaps endogenously so (ex: native flight from immigrants (Saiz and Wachter 2011)). Instead, we assume that the distribution of relative utility

[13] The median total population share of those 1,947 tracts is $p = 0.0185\%$ and the total population of NYC Chinese was $N_g = 425,210$. Assuming a binomial distribution with p and N_g, the expected count of Chinese in the median tract is $p * N_g = 78.7$ and the probability of observing 1 or fewer Chinese is zero ($4.8 * 10^{-33}$).

[14] We also considered using the entire immigrant population—pooled across all ethnic groups—as a reference group, but again, this would not capture differences between natives and immigrants. An additional issue is that population size varies dramatically across immigrant groups and thus the largest immigrant groups would have outsize effects on the threshold. For example, Mexicans are 30% of the 2010 foreign-born population nationally, and a much larger share in some cities. Then the definition of an ethnic location might reduce to simply a Mexican/non-Mexican dichotomy for those cities. Further, even if Mexicans heavily concentrate in neighborhoods for exactly the reasons specified in spatial assimilation theory (language, networks, etc...), we would be unlikely to define many neighborhoods as Mexican because the threshold would already reflect Mexican location patterns, i.e. it's difficult to be statistically different from oneself. This is a known problem when using a reference group aggregated from individual groups and some groups are very large, see Mori, Nishikimi, and Smith (2005) for a deeper discussion.

is equal, but not necessarily over any location. In other words, natives will have some locations they strongly prefer (conditional on constraints) and group g will have some, likely different, locations they strongly prefer, but the strength of these relative preferences is equal. Specifically, we assume that the *percentiles* of each group's location quotient distribution are the same. For each group, rank the J locations in ascending order by LQ and let r be the index of the rth largest location. Denoting the natives as group $g = 0$, we assume that:

$$LQ_{gr} = LQ_{0r} \Rightarrow \frac{\theta_{gr}}{\sum_{k=1}^{J} s_k \theta_{gk}} = \frac{\theta_{0r}}{\sum_{k=1}^{J} s_k \theta_{0k}} \quad \forall r \in 1..J \tag{6}$$

We then define j as an "ethnic location" for group g if the location quotient is larger than the 99th percentile of the native distribution. These locations have housing adjusted shares of group g that are larger than almost all housing adjusted native shares. Therefore the interpretation is that these locations represent group g relative utility that is unusual even in comparison to the very high end of the native location utility distribution.[15] It's worth emphasizing that our definition of an ethnic location tests a less restrictive assumption than specialization, or even the assumption in equation 6, which we use later for counterfactual simulations. Rather than comparing the group share in tract j to a tract-specific benchmark—such as s_{gj} to s_j, or LQ_{gr} to LQ_{0r}—we instead compare all tracts in a city to the same benchmark, the 99th percentile native LQ. This leads to a different interpretation: whereas a "specialized location" ($LQ > 1$) has an unusual concentration in comparison to an expectation specific to that location, our definition of an "ethnic location" is an unusual concentration relative to *any* location in the city. Since the 99th percentile of the native LQ distribution is always higher than 1, this higher threshold leads to significantly fewer ethnic locations than specialization.

To help illustrate our method, in panel A of Figure 1 we plot the location quotient for each census tract in 2010 NYC against the cumulative group population share, for five immigrant groups varying in size and the native population. The nearly vertical, thick line in the leftmost area of the graph shows the native population. For the 4661 tracts in NYC, the location quotient for natives varies from just above zero to 1.52, with the 99th percentile tract having a location quotient of 1.43. For legibility, we start the plot for the immigrant groups at tracts with a location quotient larger than this threshold of 1.43 and cut each series at the 0.95 cumulative population share.[16] If the null hypothesis in equation 6 holds, then we would expect to find roughly 1% of all locations in a city are ethnic locations. However, the figure shows that tracts with a location quotient greater than the threshold account for between 70 and 80% of the population for all five groups. By contrast, only 1.2% of the native population is in tracts with a location quotient above the threshold.

Panel A of Figure 1 also shows that the share of tracts exceeding a location quotient of one is very high. In fact, the native population share in tracts with an LQ less than or equal to one is 35%, as indicated with a point on the native series. Natives, of course, are not a homogeneous population, and it's possible that by aggregating heterogeneous groups into a single "native" classification we are masking location concentrations

[15] We considered using the maximum native LQ as the threshold but worried that measurement error in the count of natives and the tract spatial boundaries (see data discussion) would make the maximum more sensitive to outliers. Generally, any very high percentile could represent unusually high relative utility, and thus the choice of a specific threshold is somewhat arbitrary. The ad-hoc nature of thresholds is a general issue for classification, as discussed recently by (De Bellefon, Combes, Duranton, Gobillon, and Gorin 2020), who use the 95th percentile of building density for classifying a location as urban or not.

[16] The curves for some of these groups have very long tails with some tracts having an LQ greater than 30. The LQ of a tract does not depend on group size; in panel A of Figure 1 the LQ distributions for Jamaica and China generally stochastically dominate the other groups but Italy and Canada are the smallest groups, and the Dominican Republic is the largest. However, comparing the very largest LQ values across groups is not that informative because the maximum possible value does depend on group size. This occurs when the entire tract population is from the ethnic group, and thus the maximum LQ for any tract is simply the inverse of the city-level share: N/N_g.

that are just as pronounced as those of immigrant groups. Race is arguably the most studied grouping in the literature on segregation (see the survey of U.S. patterns in Logan (2013)), and so we explore this idea by disaggregating natives into large racial groups and plotting their LQ distributions. In panel B of Figure 1 we plot the LQ distributions for the largest native racial groups in 2010 New York, further disaggregating White into Hispanic and non-Hispanic origin.[17] We also re-plot the series for Italy and China from panel A for comparison. The LQ distribution for White Non-Hispanic—the largest group—is fairly similar to the overall native distribution, but the other groups have much wider LQ distributions. In particular, Black natives have tracts with much higher concentrations than the overall native distribution. This fact shows a caveat of our approach: while we seek to measure concentration above and beyond that of natives, there are large native groups with significantly higher concentrations than the average level obtained from pooling across native groups. One could try using a particularly concentrated native group as the reference group, but the size and characteristics of any such group would vary dramatically across cities, making the comparison of ethnic neighborhoods across cities much more difficult. Further, note that even the concentration of Black natives in panel B is still considerably smaller than that of the ethnic groups in panel A. The greater concentration of most ethnic groups in comparison to native racial groups holds across many cities.[18] Therefore we use natives as our reference group in order to capture some of the factors specific to immigrants, while recognizing that this benchmark understates the true concentration levels of some large native groups.

2.3 Implementation and Significance Testing

We define the 99th percentile native LQ as the location with a rank r greater than or equal to $0.99*J$, where J is the count of locations in the city. Let l^* be the index of this location and $\bar{LQ}$ be the corresponding location quotient so that $\bar{LQ} = s_{o,l*}/s_{l*}$. Then, to test whether a particular location j is an ethnic location we need to know whether the group count n_{gj} is large enough to conclude that $s_{gj} > \bar{LQ} * s_j$. Our null hypothesis is that the location count comes from a binomial distribution with probability $\bar{s}_{gj} = \bar{LQ} * s_j$ and N_g trials:

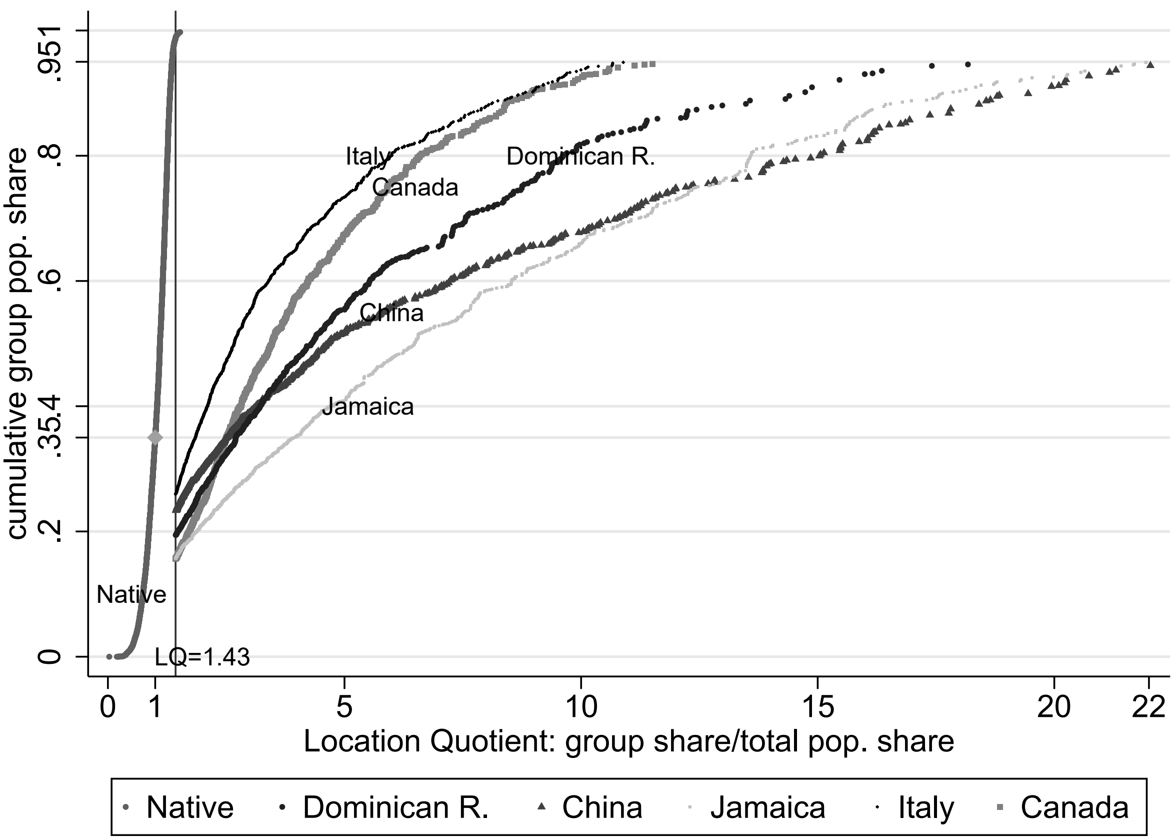
$$Pr(X \leq n_{gj}) = \sum_{i=0}^{n_{gj}} \binom{N_g}{i} (\bar{s}_{gj})^i (1 - \bar{s}_{gj})^{N_g - i} \tag{7}$$

We can use this method to get a p-value for each location and then define ethnic locations as those with a p-value less than a chosen significance level (ex: 5%). However, an important issue emphasized by Billings and Johnson (2012b) is that even if the null hypothesis is true for all locations, evaluating a large number of locations at a standard level of significance can lead to many false rejections, or type 1 errors. For example, New York City had 4661 census tracts in 2010. Even if the true preferences and constraints of an ethnic group are identical to those of the natives, it's highly likely that some of the many locations randomly receive high counts due to individual shocks (ε_{igjh}), causing us to (falsely) reject the null at $\alpha = 0.05$. To correct for this issue, we follow Billings and Johnson (2012b) and use an adjusted significance level that limits the probability of making *one or more* false rejections under the null hypothesis to 5%. This error rate is known as the family wise error rate (FWER) because errors are limited for the entire category (family) of

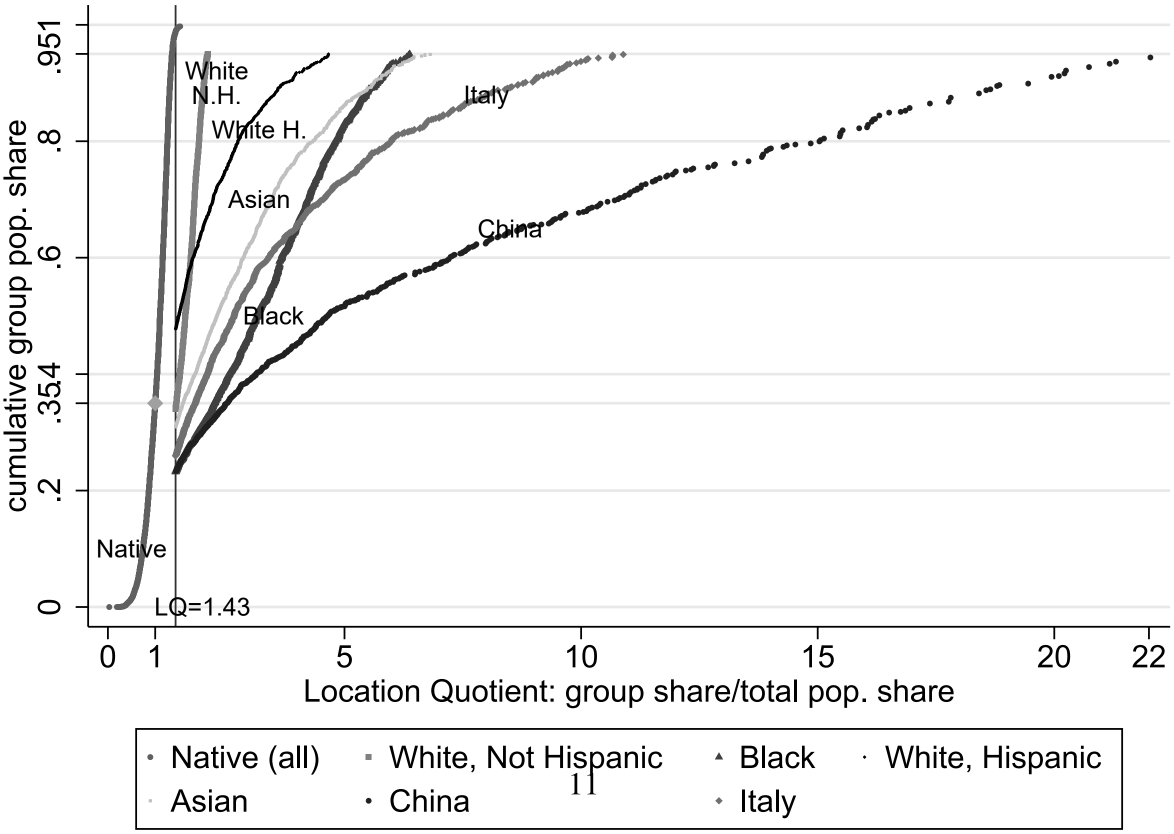
[17]The 2010 census allows respondents to choose more than one race. The counts we use for the four racial groups in panel B are for people indicating a single race (and Hispanic origin) and account for 90% of the New York City CBSA's 2010 native population.

[18]We also calculated the average location quotient for each group in 2010 by weighting the LQ of every tract by its share of the national group population. This yields the location quotient for the average member of each group. For large racial groups in the native population, we find: all natives (1.00), White not Hispanic (1.21), Black (2.87), White Hispanic (1.88), Asian (2.54). For large foreign-born groups we find: Mexico (3.07), China (6.51), Philippines (5.21), India (6.57), Vietnam (7.18), El Salvador (7.20), Korea (7.20), Cuba (6.07), Dominican Republic (8.68), Canada (4.08).

Figure 1: Cumulative Population Share by Tract Location Quotient, NYC 2010



(a) Native and Foreign Born



hypotheses, rather than for a single test. In our case the family is every location in a city, for a given ethnic group. For example, if we were to assign ethnic tracts to a city where the null hypothesis was true for every tract, using this adjusted p-value there would only be a 5% chance of making any false rejections, or 95% of the time we would correctly conclude the city did not have a single ethnic tract.

To find this adjusted p-value we use the same simulation procedure as Billings and Johnson. The basic idea is to generate random tract counts in a city for a given group under the null hypothesis and calculate p-values for every tract. The minimum p-value across all tracts is the significance level required to avoid any false rejection. This minimum p-value represents one replication at the family level and so repeating this process for many replications (we use $50,000$) yields a distribution of family-level p-values. The 5th percentile of this distribution is an adjusted p-value that limits the FWER to 5%.

In order to calculate the adjusted p-value we need a null distribution across locations for the group g population. Any distribution that is consistent with assumption 6 is sufficient, and so it's simplest to use the native distribution. However, it's worth emphasizing that assumption 6 is not equivalent to assuming natives and group g have equal location shares. For example, consider two locations k and l that have the same total population shares, $s_k = s_l$, but different native shares $s_{0k} \neq s_{0l}$. If we swap the native shares for k and l the percentiles of the native location quotient distribution are still the same. Any permutation of native shares among locations with the same total population share yields the same percentiles, and thus the native share distribution is just one realization from the set of distributions satisfying assumption 6. If housing capacity was equal in all locations then equation 6 simply implies the percentile location shares are equal, consistent with any permutation of native shares across locations.

Our null hypothesis is that no location quotient LQ_{gj} exceeds the 99th percentile of the observed native distribution, $\bar{LQ}$. Therefore, we want to choose an adjusted p-value that limits the probability that if we drew people randomly from the native distribution, one or more of the tracts in the first 99 percentiles would randomly exceed this observed $\bar{LQ}$. Let r be the rank of this 99th percentile tract. We first draw N_g people from the native population across locations *with* replacement. We then calculate the minimum p-value from testing the null hypothesis that each simulated share s_{gj} is less than or equal to $\bar{LQ} * s_j$—using the same binomial test as above—but only among the first r tracts. This gives us the minimum significance level such that no tract below the 99th percentile exceeds the threshold location quotient for that replication. We then repeat the procedure $50,000$ times and use the 5th percentile as our adjusted p-value. To implement draws from the set of native shares we use the multinomial distribution; see Appendix section 7.1 for further detail.

Lastly, while our model describes the choice of households, our data only provides counts of individuals by ethnic group and nativity. Inference in our method (p-values) depends on the number of residential choices, but using individuals to define ethnic tracts likely overstates the number of independent choices observed since most individuals live in multi-person households. Using the total tract population divided by the count of households in each tract as an approximation of household size, the median household size across all cities and years is 2.76. If we restrict to census tracts with high immigrant fractions (10%, 20%, 30%), the median household size is also between 2.8 and 3. As a conservative approximation, we assume the household size for all groups (including natives) is two people, and therefore divide all tract counts by two and use this as a measure of households. The location quotient is defined in terms of shares, so dividing the tract population and total population by a constant has no effect on the LQ for any tract. However, this adjustment does affect p-values, generally raising the required count for ethnic tract status in lower population tracts.

2.4 Clustering Tracts into Neighborhoods

The map in panel A of Figure 2 shows that many of the locations with the highest Chinese population are clustered into groups of contiguous tracts. Further, tracts with the largest concentration of Chinese tend to have many Chinese tracts in the vicinity whose Chinese populations decrease with distance, seen visually as a roughly monotonic fade in shade intensity away from the brightest colored dots. Given these patterns, we define "ethnic neighborhoods" as a unique cluster of spatially contiguous ethnic tracts. Two tracts are spatially contiguous if they share any portion of their borders, a contiguity measure often referred to as "Queen's" after the movement of the chess piece. We then group together clusters of overlapping neighbors such that if tract A and tract B are in the same neighborhood, then it is possible to walk from A to B without stepping foot in a non-ethnic tract.

We provide interactive maps showing ethnic tracts and neighborhoods for many groups, cities, and years on our mapping website: https://nathanschiff.shinyapps.io/ee_maps_deployed/. In panel B of Figure 2 we show the results from applying the grouping algorithm to the 2010 New York City Chinese population, centering the map on neighborhoods with the largest Chinese populations. We shade the polygons from white (zero Chinese) to black (7649 Chinese) and use a square root coloring scale in order to show differences in population across tracts with lower counts.[19] In panel B of Figure 2, the large Chinese clusters of tracts shown with red borders match well with some of the known Chinese neighborhoods, labeled in panel A. The map also shows a number of singleton Chinese tracts, as well as cases where two different neighborhoods are only separated by a few non-Chinese tracts. In their paper on industry clusters, Mori and Smith (2013) use a more sophisticated method to group together all contiguous regions with high concentrations of an industry that form a convex set. If applied to our context, this method would add tracts on the shortest path between two ethnic tracts, and likely merge some of the neighborhoods we have identified. In Logan, Zhang, and Alba (2002), the authors add locations adjacent to their identified ethnic tracts—what they call "edge tracts"—to the ethnic neighborhoods, which also smooths out some of the contiguity issues. While these methods are logical ways of defining clusters, our method has the property that each tract in a cluster individually satisfies the ethnic tract threshold. This makes it simpler to study longitudinal questions, such as how the characteristics of a specific tract affect its likelihood of joining or leaving an ethnic neighborhood. Further, several papers have shown that neighborhoods can sometimes have sharply defined borders[20] and therefore we don't wish to mistakenly smooth away true spatial discontinuities.

3 Characteristics of Tracts and Neighborhoods

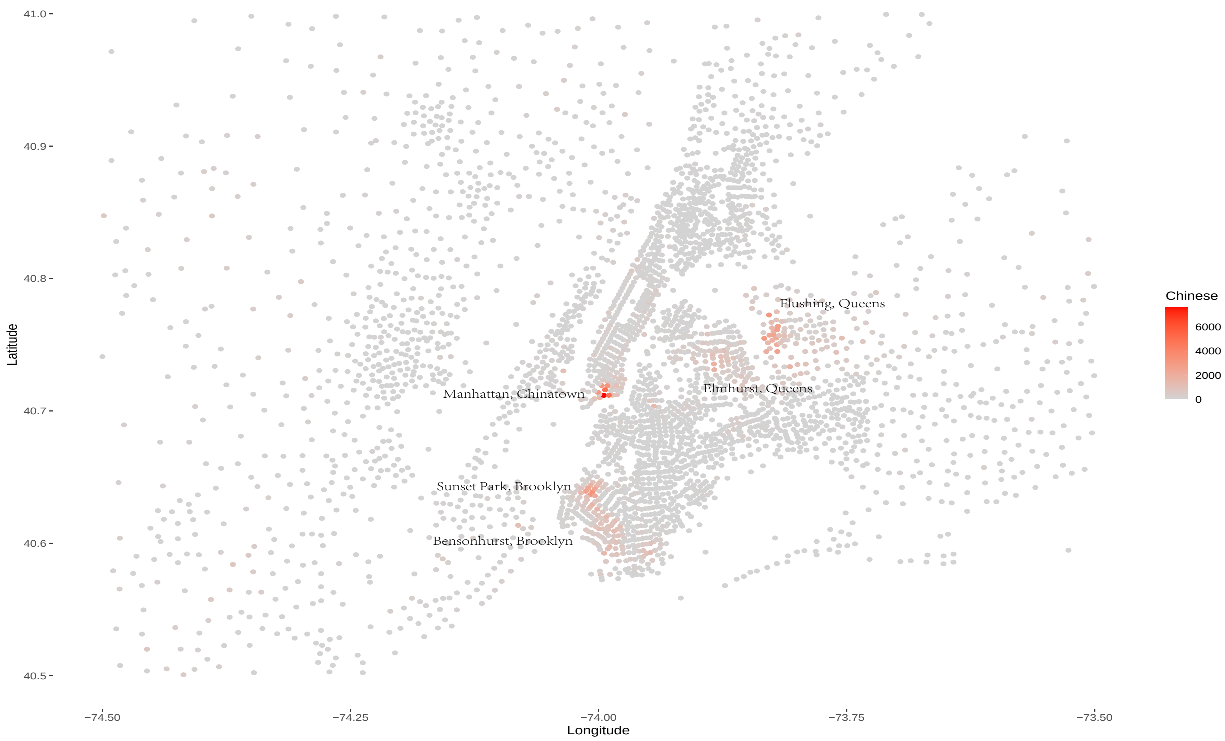
3.1 Data

We use data from the 1970, 1980, 1990, and 2000 US censuses, as well as averages from five consecutive American Community Surveys (2006-2010), which we will refer to as 2010 data. All of this data is at the census tract level of geography, with variables standardized by the National Historical Geographic Information System at the Minnesota Population Center (Manson, Schroeder, Riper, and Ruggles 2018). We have converted this data to constant boundary census tracts, using the 2010 boundaries, so that the data can be

[19] If we used a linear scale, as in panel A, only the neighborhood tracts would have any noticeable shading. The square root scale is often used for count data since it allows for zeros and is variance stabilizing, under the assumption the counts are distributed Poisson.

[20] Aaronson, Hartley, and Mazumder (2017) show that "redlining" maps led to boundary discontinuities in racial location patterns that persist to this day. Ananat (2011) argues that railroad tracks can separate neighborhoods while Cutler and Glaeser (1997) make a similar argument for topographical features.

Figure 2: New York City Chinese Tracts, 2010



(a) Census Tract Centroids

(b) Census Tracts Polygons

*Notes:*Panel A plots the centroids of NYC census tracts, shading each dot by count of Chinese born. Panel B plots a subset of NYC census tracts, shading each tract by count of Chinese, using a square root scale. The red borders show clusters of contiguous Chinese tracts defined using our algorithm.

considered a panel of census tracts.[21] For each census tract in each year, we have data on housing and demographic characteristics, including population counts by birth country. The conversion to constant boundary tracts often results in fractional values for count variables. In these cases, we take the largest integer value less than or equal to the fraction (often called the "floor" or "int" function). In 1970 and 1980 the census question on country of birth only allowed for a limited set of country responses, but the response set expanded significantly from 1990 onwards. As a result, the length of our panel differs by group. All of our data, including the spatial boundary files and crosswalks, is publicly available.

Most of our variables, including the counts by birth country and nativity, are summary level data from the long-form of the census, which are estimates from a sample rather than full population counts. This introduces measurement error into our immigrant group counts, and therefore some of the tracts we observe with zero individuals from a given group may actually have residents from that group. Similarly, we also observe some tracts where the count of natives exceeds the total tract population, where the total tract population is measured using the full population counts. We drop any cases where the count of natives exceeds 1.1 times the total tract population.[22] An additional source of measurement error is the conversion to constant boundary census tracts. Since we focus on unusually large concentrations, it's unlikely that any of this measurement error would affect the classification of many ethnic tracts, nor change any of the general patterns we find. Nonetheless, it is likely that a few tracts on the margin of the LQ threshold could change status with more accurate population counts. Of course, these tracts would also change if we changed the native threshold percentile, and thus our focus is on general patterns rather than whether a specific location is an ethnic tract. Lastly, in some of our analysis comparing results over time, we see a trend break at 2010 where there is a fairly consistent pattern from 1970-2000 and then a change in the 2010 statistic. This discrepancy could result from the conversion to constant boundary tracts—the 1970-2000 are converted while 2010 are the original boundaries—or from differences between the decennial census data (1970-2000) and the American Community Survey five year average we use for 2010. Therefore we limit our time series analysis to robust patterns that are present across many decades, rather than specific changes between any two decades.

3.2 Comparison with other Neighborhood Definitions: Language Use

Our method for defining ethnic neighborhoods is more complicated than some of the simpler methods in the literature, such as those based on the ethnic percentage of a region. Therefore, it's important to show that this additional computation also yields additional value. While there is no standardized set of characteristics associated with ethnic neighborhoods, one important feature commonly discussed is that residents of ethnic neighborhoods are more likely to use the origin country language (Bauer, Epstein, and Gang 2005, Chiswick and Miller 1996). In this section we provide evidence showing that our method provides additional power in predicting language use beyond controlling for the ethnic population or defining neighborhoods based on ethnic percentage cutoffs. We also perform a similar analysis for ethnic population growth in section 5.2. In section 3.4 we show that ethnic neighborhoods defined with our method follow a power law, while tract-based measures do not.

Below, we first investigate whether ethnic tract status and neighborhood population size can predict language behavior, conditional on ethnic population. We then compare our neighborhood definition directly

[21] We did this using the census boundary crosswalks provided by the S4 Institute at Brown University. Essentially 2010 census block boundaries were overlayed on the original boundaries and then populations were reallocated assuming a uniform distribution within blocks (or sometimes tracts), correcting for the amount of surface area covered by water—see Logan, Xu, and Stults (2012) for details.

[22] Since there must be cases where the native population is underestimated, we were concerned that dropping all cases where the native population exceeded the total population would be an asymmetric adjustment. Therefore we chose 1.1 as an approximation allowing for 10% error.

to a simpler definition of an ethnic tract based on cutoff values in ethnic percentage. We show that in a regression of language use on ethnic tracts defined by ethnic percentage, adding the neighborhood population from our definition always improves the fit.

Our null hypothesis, following Bauer, Epstein, and Gang (2005) and others, is that members of an ethnic group living in an ethnic neighborhood will be more likely to speak the birth country language at home than members of the group not living in a neighborhood. To test this, we use the count of people speaking a specific language at home from the 2010 census, available at the tract level. We examine language spoken with two specifications for the count of people in CBSA c, in census tract j, speaking a given language. The first is:

$$\begin{aligned} spkEthLang_{cj} =& \beta_1 * ethPop_{cj} + \beta_2 * nonEthPop_{cj} + \beta_3 * ethtract_{cj} \\ &+ \beta_4 * (ethPop_{cj} \times ethtract_{cj}) + \beta_5 * (nonEthPop_{cj} \times ethtract_{cj}) + \mu_c + \varepsilon_{jc} \end{aligned} \tag{8}$$

The variables "ethPop" and "nonEthPop" represent the ethnic population for the corresponding language (e.g., the count born in China for the dependent variable Chinese speakers), and the interactions of these variables with the ethnic tract indicator allow us to test whether language behavior for the two populations differs in and out of ethnic tracts. We also include CBSA fixed effects, μ_c. The second specification takes the same form, but we replace the ethnic tract variable and interactions—the variables corresponding to coefficients β_3, β_4, and β_5—with the sum of the ethnic population from immediately adjacent tracts and the size of the ethnic population in the neighborhood containing tract j. This allows us to see whether neighborhood population can help to explain the language behavior of a component tract, even conditional on the tract population and the immediately adjacent tract populations. We restrict the sample to CBSAs with at least 1000 members from the ethnic group, and census tracts in those CBSAs with at least one member of the ethnic group.

In the top panel of Table 1 we show the results for four of the ethnic groups with a unique language tracked by the 2010 census. This makes the analysis simpler than for ethnic groups that share a language with many other groups, such as Mexicans. The first coefficient in this table shows that each additional person born in China increases the count of Chinese speakers by 1.2 people. Note that it is not unusual that this coefficient is greater than one because there may be many speakers of a language who were not born in the corresponding country (ex: US-born children of Chinese immigrants). Additionally, the language spoken variable is a summary statistic with measurement error that may not exactly match that of the birth country variable. The coefficient on "non-ethnic population" is also positive and significant, but much smaller, implying that for every 1000 people in a tract who were not born in China, there will be a single additional Chinese speaker. The main effect for "ethnic tract" is insignificant in this column, but the ethnic population interaction implies that an additional Chinese resident in a Chinese ethnic tract is associated with an additional 1.3 Chinese speakers, or 0.06 more than in non-Chinese tracts. Below the coefficients we list the mean of the dependent variable ("dep-var mean"), which is 121 Chinese speakers in column one. The row "nb-var mean" lists the mean of the population in an ethnic tract for odd-numbered columns, and for even-numbered columns it lists the mean in the containing neighborhood. The average Chinese ethnic tract in column 1 has 242 Chinese, and thus the average Chinese tract would have an additional 14.7 Chinese speakers more than a non-Chinese tract with the same Chinese population, or about 12% of the mean (14.7/121). Thus, our ethnic tract indicator provides additional information beyond the ethnic population count alone.

The results in column two show that an additional 1000 people in the containing neighborhood is associated with an additional 1.028 Chinese speakers. The mean neighborhood population is $10,810$ Chinese, which implies that for the average neighborhood, each tract is predicted to have an additional 11.11 Chinese speakers. Note that this coefficient is conditional on both the tract populations and the Chinese population in immediately adjacent tracts, whose coefficient implies that each additional 1000 people in adjacent tracts is associated with 6.4 additional Chinese speakers. Some of the adjacent tracts are part of the containing neighborhood, and thus the neighborhood coefficient is identified from neighborhoods that extend beyond the immediately adjacent tracts, or do not include all of the adjacent tracts. Intuitively, neighborhoods can have complicated spatial arrangements and thus our neighborhood measure captures more than just the effect of immediately adjacent tracts. For Italians and Koreans, we also find positive and significant interactions between ethnic tract and ethnic population, but the coefficient is insignificant for Vietnamese. For Italians and Vietnamese, the neighborhood population size is significant, but not for Koreans.

Table 1: Ethnic Tracts, Neighborhoods, and Language Spoken at Home

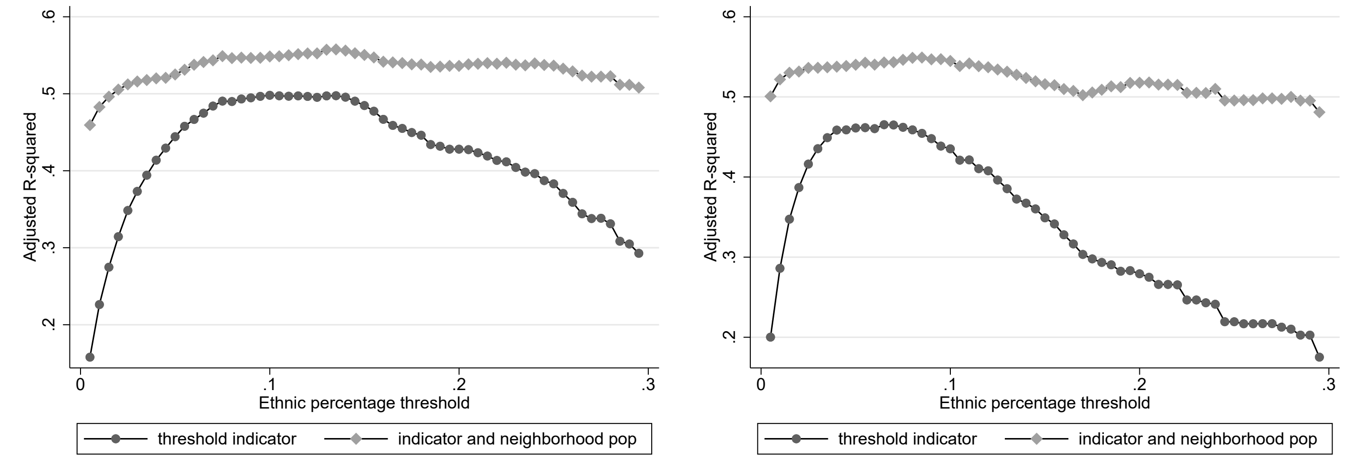
	(1) China	(2) China	(3) Italy	(4) Italy	(5) Korea	(6) Korea	(7) Vietnam	(8) Vietnam
ethnic population	1.233***	1.231***	1.283***	1.310***	1.050***	1.114***	1.123***	1.078***
	(0.026)	(0.008)	(0.034)	(0.065)	(0.033)	(0.016)	(0.023)	(0.026)
non-ethnic population	0.001***	0.002***	0.001***	0.001***	0.000	0.000	0.001***	0.002***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ethnic tract	-8.062		-9.872***		-8.093***		-20.852**	
	(5.676)		(2.030)		(2.665)		(7.960)	
eth. pop. X eth. tract	0.061**		0.215***		0.100***		-0.009	
	(0.030)		(0.024)		(0.027)		(0.025)	
non-eth. pop. X eth. tract	0.001*		-0.001		0.000		0.004***	
	(0.001)		(0.000)		(0.000)		(0.001)	
adj. eth. pop. 000's		6.384*		50.441***		9.151***		3.797**
		(3.477)		(6.222)		(1.868)		(1.476)
neigh. pop. 000's		1.028***		1.445***		-0.042		0.357**
		(0.392)		(0.380)		(0.298)		(0.139)
Observations	20345	20345	9844	9844	15550	15550	13474	13474
Clusters	130	130	45	45	105	105	103	103
dep-var mean	121	121	52	52	65	65	87	87
nb-var mean	242.31	10.81	69.89	0.88	148.47	4.97	173.05	6.51
Adj. R2	0.96	0.96	0.77	0.78	0.96	0.96	0.92	0.92

Notes: Dependent variable is count of people speaking the language of the column header. The variable "adj. eth. pop." is the sum of the ethnic population from adjacent tracts and "neigh. pop." is the ethnic population of the containing neighborhood, both measured in thousands. The table row "nb-var mean" reports the mean of the ethnic population in ethnic tracts for odd columns; for even columns this is the mean of the containing neighborhood population. The sample is 2010 census tracts with at least one member of the ethnic group, in cities with at least 1000 people in the ethnic group. All specifications include CBSA fixed effects and standard errors are clustered by CBSA.

A related area of interest is whether immigrants in ethnic neighborhoods speak English at a lower level of proficiency than immigrants in non-ethnic neighborhoods. In Appendix section 7.2 we use a similar analysis to the above and find that residents in ethnic neighborhoods have a significantly higher likelihood of reporting they "speak English less than very well."

The above results show that our neighborhood measure provides additional information that can predict language use, beyond the ethnic population of a tract. However, this is not a direct comparison to defining ethnic neighborhoods with a simpler method. Therefore, we now compare our measure to tract-level neighborhoods based on cutoff values in ethnic percentage. Specifically, a tract j is defined as an ethnic neighborhood if $ethpop_j/totpop_j \geq k$, for some cutoff value k. We then regress the count of people speaking

Figure 3: Comparing Cutoff-based Measures to LQ-defined Neighborhoods



(a) Speaking Chinese

(b) Speaking Korean

Notes: Each point on a series shows the adjusted R-squared from regressing the count of people speaking Chinese (A) or Korean (B) in a tract on an indicator for whether the tract ethnic percentage ($ethpop_j/totpop_j$) is above the cutoff value shown on the horizontal axis. In the lower series the regressions only include the indicator while in the upper series we include both the indicator and the ethnic population size of the containing neighborhood. For tracts outside of ethnic neighborhoods, the neighborhood population is zero. We exclude ethnic neighborhoods with a single tract from all regressions so that neighborhood population is never equal to the tract ethnic population.

the origin country language on this tract-level neighborhood indicator and record the adjusted R-squared. We then run a second regression where we also include the neighborhood population from our definition ("LQ-defined neighborhoods"), and compare the resulting adjusted R-squared to the first regression. The sample for both regressions excludes singleton neighborhoods defined with our method so that the neighborhood population is never simply equal to the tract ethnic population. We estimated these models for Chinese speakers/Chinese tracts and Korean speakers/Korean tracts, but we find similar results for other groups. We performed this exercise for a range of cutoff values, $k \in [0.005, 0.3]$ and plot the results in Figure 3. In each panel there is an inverse U-shaped relationship between adjusted R-squared and the cutoff value, with the cutoff value that maximizes adjusted R-squared varying by group. Importantly, for all cutoff values in both panels, the regressions including neighborhood population yield a significantly higher adjusted R-squared.

3.3 Tract Characteristics

We now turn to describing characteristics of ethnic tracts (this section) and ethnic neighborhoods (next section) defined with our method. One strand of the literature on U.S. immigration suggests that geographic concentration of recent arrivals may be less important than in the past due to changes in the skill profile of immigrants, an increased diversity of settlement locations, as well as technological improvements in transportation and communications (see discussion in (Gold 2015), page 10). Portes and Rumbaut note that "professionals and technicians...seldom cluster in highly visible ethnic communities" (Portes and Rumbaut 2014, p. 36). Since the share of foreign-born workers in the U.S. with a college degree or working in high-skill occupations has increased noticeably over time (Budiman, Tamir, Mora, and Noe-Bustamante 2020, Bennett 2020), this change could lead to smaller or fewer ethnic neighborhoods. Hirschman and Massey (2008) write that immigrants arriving in the 1990s and later were more likely to locate in smaller cities and towns than previous cohorts, another pattern which could reduce the importance of existing ethnic neighborhoods and lower overall sizes. On the other hand, Portes and Rumbaut emphasize that U.S. immigration patterns show a strong spatial persistence such that "after several generations particular nationalities continue to be associated with specific patches of national territory"(Portes and Rumbaut 2014, p. 107), although they are referring to regions, rather than neighborhoods.

To evaluate whether ethnic neighborhoods have decreased in size or prevalence, we calculate the percentage of the foreign-born population living in ethnic tracts, along with characteristics of those ethnic tracts, for each census year, and show the results in Table 2. We can only define ethnic tracts for the portion of the foreign-born population for which the US census lists the specific birth country. There is a separate variable denoting the count of foreign-born individuals (summed over all countries) in a tract, which we use to calculate the total foreign-born population. We further limit the sample to those cities with at least 100 tracts, which ensures we can measure the 99th percentile of the native population with minimal error. This sample with listed birth countries represents 71 to 91 percent of the foreign-born in cities with at least 100 tracts, and 65 to 80 percent of the foreign-born population across all cities; see table notes for percentages by year. We estimate the percentage of the foreign-born living in ethnic tracts in a given year as the sum of the ethnic population in ethnic tracts across all classified birth countries, divided by the total foreign population living in cities with at least 100 census tracts. While there is a debate in the literature about the importance of residential location for immigrant populations, the results in Table 2 show that the percentage of the foreign-born living in ethnic tracts has increased significantly over time from 43% in 1970 to 67% in 2010. In columns 2-5 we show characteristics of ethnic tracts for the average foreign-born resident living in an ethnic tract. To do so, we weight the characteristics of each ethnic tract by the fraction of the total ethnic population living in ethnic tracts, represented by that tract.[23] Column 2 shows that the size of the co-ethnic population in ethnic tracts (e.g., the count of China-born residents in a Chinese ethnic tract) has also increased from 384 people in 1970 to 683 people in 2010. As mentioned at the end of section 3.1, the 2010 value for this column, as well as some of the others, breaks with the trend of a monotonically increasing tract population. However, again, we don't want to emphasize this change since it could reflect differences in the 2010 data collection, but rather highlight the overall pattern of significant growth since 1970. Column 3 shows that the location quotient for the average ethnic tract has stayed fairly constant over time, with the share of the ethnic population in an ethnic tract between ten and eleven times larger than the overall population share. The ethnic fraction of the tract—the ethnic population divided by the total population of the tract—has increased dramatically, but not monotonically, from just under 9% to over 13% in 2010.

Another finding in the U.S. immigration literature is that some ethnic neighborhoods can now be found outside of central cities in more suburban locations (Logan, Zhang, and Alba 2002, Gold 2015). However, it's unclear whether this is a widespread phenomenon. We examine this idea with two measures of centrality: simple population density (total population divided by land area in square kilometers) and distance in kilometers to the central business district (CBD), as defined by the 1982 Census of Retail Trade and implemented by Fee and Hartley (2013). While distance to the CBD is a simpler measure of centrality, CBSAs are quite large and often have multiple sub-cities or employment centers, and thus population density may better capture this polycentricity. As it turns out, we find fairly similar trends across the two measures. In column 5 we show the average population density for ethnic tracts in every year and find that this decreased from 9329 people per square kilometer in 1970 to 5590 people in 2010, a decrease of 40%. In column 6 we show

[23] Let G_t be the set of foreign countries for which the US census lists birth country in year t. Let J_{gt} denote the set of ethnic tracts across all cities (with at least 100 tracts) for group g in year t. Finally, let $ethpop_{gjt}$ be the population born in country g living in tract j in year t. Then the value of tract characteristic x for the average foreign-born resident living in an ethnic tract is $\sum_{g \in G_t} \sum_{j \in J_{gt}} \frac{ethpop_{gjt}}{N_t} * x_{jt}$, where $N_t = \sum_{g \in G_t} \sum_{j \in J_{gt}} ethpop_{gjt}$ is the total population across all groups living in ethnic tracts in year t.

the tract population density for the average native—we weight each tract's population density by its share of the national native population—and we find a similar decrease of about 40%. In column 7 we find that the average distance to the CBD increased from 14km in 1970 to 21.8km in 2010, a 56% increase. The distance to the CBD for natives increased by 33% over the same period. This is somewhat less than the amount for ethnic tracts and for any pair of years the percentage change is larger for ethnic tracts, meaning the larger increase does not depend on the year range chosen. These statistics show that ethnic neighborhoods did greatly decentralize, consistent with the literature. However, we find the same density pattern for natives and only a modestly larger increase in distance to the CBD for ethnic tracts. Therefore, additional evidence would be needed to argue that the decentralization of ethnic neighborhoods reflects factors specific to immigrants, rather than a general decentralization of the entire US population. Note that the average population density in ethnic tracts remains more than twice as large as the tract density for the average native, and the distance to the CBD is still 20% less than for natives.

Table 2: Ethnic Tracts: Population Statistics

Year	Pct. in E. Tracts	Av. Ethnic Pop.	Average LQ	Av. Ethnic Fraction	Av. E. Tract Density	Av. Native Density	Av. E. Tract CBD Dist.	Av. Native CBD Dist.
1970	43.1%	384	10.7	0.088	9329	3969	14.0	20.5
1980	49.6%	616	10.1	0.142	7412	2897	16.1	23.1
1990	57.0%	641	10.7	0.133	7498	2580	17.9	25.4
2000	60.1%	825	9.0	0.159	6979	2522	19.7	26.5
2010	67.2%	683	11.3	0.131	5590	2390	21.8	27.3

Notes: The first column shows the percentage of the foreign-born population living in ethnic tracts. The second column is the average co-ethnic population in an ethnic tract and the third column is the average location quotient. The fourth column is the average fraction of an ethnic tract's total population comprised of the ethnic population: ethnic population divided by total population. The fifth column shows the average population density—total population divided by land area in square kilometers—of an ethnic tract. The sixth column shows the population density of a tract for the average native (born in US). Columns seven and eight show the average distance in kilometers from a tract to the city central business district (CBD) for ethnic tracts and natives. Ethnic tract statistics for a given year are calculated using the foreign-born population for which the census lists the birth country, and only for cities with at least 100 census tracts. In cities with 100 tracts, this sample accounted for the following percentages of the total foreign-born population: 75% (1970), 71% (1980), 89% (1990), 86% (2000), 91% (2010). As a percentage of the total foreign-born population across all cities (including cities with fewer than 100 tracts), the corresponding percentages are: 70% (1970), 65% (1980), 79% (1990), 76% (2000), 80% (2010). See text for weighting and calculation details.

In Table 3 we compare characteristics of ethnic tracts to other tracts where an ethnic group lives. For most ethnic groups, every city has a large number of locations without a single person from that group. By restricting the sample to only locations with positive counts, we can separate the unique characteristics of a group's ethnic tracts from the more general differences in the location choices of the group relative to all tracts in the city. In order to calculate the average effect for the foreign-born population in a given year, we pool the data across groups and weight each group-tract observation by its share of the total foreign-born population in that year.[24] Specifically, we run regressions from the following pooled specification, separately by year:

$$y_{gcj} = \beta * ethtract_{gcj} + \mu_{gc} + \varepsilon_{gcj} \tag{9}$$

In this specification $ethtract_{gcj}$ indicates whether tract j in city c is an ethnic tract for group g. Note that the same tract can be an ethnic tract for multiple groups. We include city-group fixed effects, μ_{gc}, to capture

[24]The weight for group-tract observation gj is $ethpop_{gj}/N_t$, where $N_t = \sum_{g\in G}\sum_{j\in J} ethpop_{gj}$ is the total foreign-born population in year t. Similar to Table 2, this weighting attempts to back out the coefficient we would obtain if we had individual-level data, rather than tract-level.

average differences across city-groups in a given year, and cluster standard errors at the city-group level. The coefficient β can thus be interpreted as the difference in the outcome variable for the average foreign-born resident in an ethnic tract, compared to the average foreign-born resident in a non-ethnic tract. We run this specification for six housing and income variables across the five census years, thus each cell in Table 3 represents the estimate of β (and standard error) from a separate regression.

In column 1 of Table 3 we find that people living in ethnic tracts are much more likely to be renters than people in non-ethnic tracts where the ethnic group also lives, but this difference has decreased substantially over time from 12.7% in 1970 to 8.1% in 2010. Note that this variable, and all others in Table 3, refers to characteristics of the entire population of the tract; there is no public data for outcomes by birth country at the tract level. Cutler, Glaeser, and Vigdor (2008a) suggest that immigrant segregation can be partly explained by differences in transportation mode usage. Other research has shown that new immigrants use public transit at much higher rates than native-born Americans, but that this transit use declines with the duration of stay (Chatman and Klein 2009, Blumenberg and Evans 2010). In column 2 we show the fraction of the tract commuting to work without a car and find that this commute mode is much more likely in ethnic tracts but has also decreased considerably over time. We find a similar pattern for the fraction of the tract living in housing built more than 30 years earlier (relative to the census year). In columns 4-6 we examine tract median household income (thousands of dollars), median household rent (hundreds), and median housing value (hundred thousands), with all dollar values converted to 2010 dollars using the CPI. Column 4 shows that median household income is much lower in ethnic tracts, but this difference has stayed fairly consistent over time, ranging between 7 and 9 thousand dollars lower. Ethnic tracts also have consistently lower housing rents, with the difference ranging between \$66 and \$83 per month and no apparent time trend. On the other hand, housing values are much lower in ethnic tracts and this difference has widened over time, from \$9,600 in 1970 to \$25,100 in 2010. The first three columns, as well as the density estimates from Table 2, suggest that ethnic tracts in 2010 are less likely to be located in dense, central city neighborhoods with older housing and a reliance on public transportation or walking than in 1970. However, these changes in location have not coincided with a convergence in household income, and residents of ethnic neighborhoods continue to live in much cheaper housing than in tracts outside the neighborhood where other members of the group live.

While Table 3 shows how the average resident of an ethnic neighborhood has fared over time, these averages mask considerable heterogeneity across groups. In Appendix Table 13 we run the same specifications, but separately by ethnic group for a large selection of groups, including natives and pooled foreign-born. Household incomes are significantly lower in ethnic tracts for some groups (for example, Mexicans and Dominicans), with a disparity that has grown over time. However, for other groups this difference has become insignificant over time (e.g., Chinese), and in some cases, ethnic tracts actually have higher median incomes (e.g., Indian tracts in 2000 and 2010). There is similar heterogeneity in the estimates for housing rent and housing values. For the other variables, the time trends are more similar across groups, with the rental percentage, fraction commuting without a car, and fraction living in housing built more than 30 years earlier declining for every group.

3.4 Neighborhood Characteristics

We start by showing the neighborhood size for the average foreign-born resident living in an ethnic neighborhood over time. As shown in Table 4, the average ethnic population of an ethnic neighborhood has increased substantially over time from 18,310 (co-ethnic) people in 1970 to 42,425 in 2010. The neighborhood size for the *median* foreign-born resident has also increased over time, but in each year it's around one tenth the size

Table 3: Ethnic Tracts: Housing, Income, and Commuting

Year	Rental Fraction	Cmt. No Car Fraction	H. Age. Gt. 30 Yrs. Frac.	Med. H.H. Income	Med. H. Rent	Med. H. Value
1970	0.127	0.094	0.151	-8.816	-0.817	-0.096
	(0.007)	(0.009)	(0.009)	(0.515)	(0.085)	(0.016)
1980	0.107	0.077	0.122	-6.972	-0.751	-0.152
	(0.007)	(0.007)	(0.009)	(0.698)	(0.112)	(0.047)
1990	0.095	0.064	0.073	-9.258	-0.833	-0.178
	(0.006)	(0.005)	(0.007)	(0.733)	(0.091)	(0.081)
2000	0.086	0.046	0.057	-8.674	-0.658	-0.188
	(0.006)	(0.006)	(0.009)	(0.684)	(0.098)	(0.044)
2010	0.081	0.029	0.070	-9.625	-0.752	-0.251
	(0.006)	(0.005)	(0.011)	(0.796)	(0.095)	(0.051)

Notes: Each cell displays the coefficient from a regression of the column's dependent variable on an ethnic tract indicator, controlling for group-city fixed effects and run separately by year (see specification in text). The standard errors shown in parentheses are clustered at the group-city level. The sample is the same as that in Table 2 but further restricted to tracts with at least one resident from the ethnic group. The regression is weighted by share of the total foreign-born population and thus each coefficient represents the difference in a given characteristic between ethnic tracts and other tracts where members of the ethnic group live, for the average foreign-born resident in a census year. The dependent variable in column 1 is the fraction of the tract living in rental housing, column 2 is the fraction commuting without a car, and column 3 is the fraction of the tract living in housing built more than 30 years ago. Column 4 is the tract's median household income in thousands, column 5 is median housing rent in hundreds, and column 6 is the median housing value in hundred thousands. The values in columns 4-6 have been converted to 2010 dollars using the CPI from the Bureau of Labor Statistics.

of the average.[25] The large disparity between mean and median stems from the presence of a small number of very large neighborhoods for every group. In fact, we now show that the distribution of neighborhood population sizes follows a power law for nearly every group, implying that the biggest neighborhoods are orders of magnitude larger than the smallest.[26]

In Figure 4 we show an example for Chinese neighborhoods in 2010. We first rank the neighborhoods by the count born in China (rank one is the largest) and then plot the rank of each neighborhood against the population, using a log scale for both axes. This type of plot is known as a rank-size plot and is a common way to show power law relationships, see (Gabaix 2016). In panel A the circular markers starting from a rank just below 2000 show that the relationship between log rank and log neighborhood population starts as a curve and then quickly turns in a straight line with a negative slope starting around a neighborhood population of 50. The line running along the linear part of this curve shows the fit line from regressing log rank on log neighborhood population across all the neighborhoods, yielding a slope of -0.78 and an R-squared of 0.98. We also perform the same exercise for ethnic tracts, as shown by the higher scatter plot (triangular markers) in panel A. While the curves are quite similar for smaller populations—many small ethnic neighborhoods are just single tracts—at larger populations the curve shoots downwards. This pattern is not characteristic of a power law, as shown by the poor match with the corresponding fit line.

One explanation for this neighborhood-level power law is that it simply reflects a power law distribution

[25] To calculate the value of a characteristic X for the median resident, we first sort all ethnic neighborhoods in ascending order by X. We then calculate a running sum of foreign-born residents for these sorted neighborhoods. We define the median value of X as X_m, where m is the neighborhood that houses the median foreign-born resident such that the population running sum of neighborhood $m-1$ is below the median and the sum for m is above or equal to the median.

[26] If a variable X follows a power law then the probability X is larger than some fixed value x is $P(X > x) = a/x^{\zeta}$, where a is a constant and the exponent ζ is the parameter of interest (Gabaix 2016). This relationship is often shown by ranking the values of X and then relating the rank of observation X_i to the value X_i using $Rank_i = a * X_i^{-\zeta}$. The famous empirical regularity Zipf's Law for cities specifies that the population rank of a city is proportional to 1/population, or that $\zeta = 1$

at the city level. If, for example, ethnic populations are distributed across cities following a power law, then randomly sampling a fixed proportion of the population in each city would also follow a power law. In panel B of Figure 4 we again plot Chinese neighborhood rank against neighborhood population, this time focusing only on neighborhoods above the 25th percentile of neighborhood population, which is roughly where the linear section of the curve starts. For each neighborhood, we also plot the corresponding Chinese population of the CBSA (gray circles). These points are always to the right of the fit line since the Chinese population of a neighborhood with rank r must always be less than the Chinese population of the entire city. The CBSA populations show as vertical lines because for each CBSA there are neighborhoods of all sizes. The largest CBSAs do have the largest neighborhoods, but they also have many small and medium sized neighborhoods. This pattern shows that the neighborhood-level power law distribution cannot be explained by the distribution of Chinese populations across cities.[27]

To more systematically examine whether the neighborhood size distributions follow power laws, we run the following specification separately for the ten largest foreign-born populations in 2010:

$$ln(nbrank_n - 1/2) = \alpha + \zeta ln(nbpop_n) + \varepsilon_n \tag{10}$$

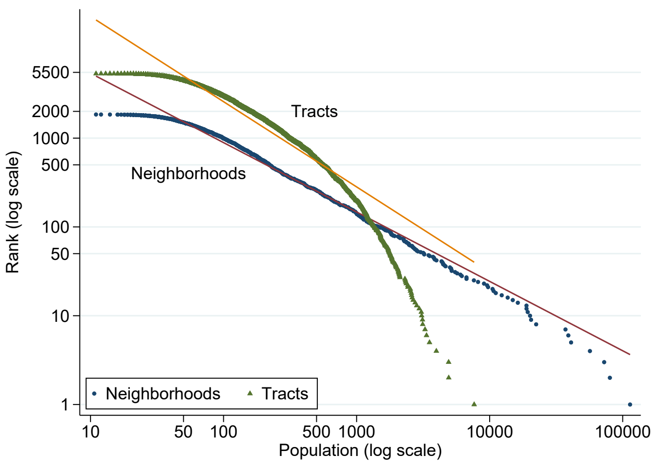
The dependent variable is the natural logarithm of neighborhood rank minus 1/2, which corrects for small sample bias in estimating a power law exponent using ζ, see (Gabaix and Ibragimov 2011) and (Rozenfeld, Rybski, Gabaix, and Makse 2011). For all ethnic groups, the rank-size pattern is similar to the plot in panel A of Figure 4, with a flat section that quickly becomes a line with a negative slope. The point where the linear section begins is often close to the 25th percentile and so we limit the regression sample to neighborhoods above the 25th percentile for each group.[28] We show the results of estimating equation 10 for the ten largest groups in 2010 in Table 5. The coefficients range from -0.7 (Mexico) to -1.39 (Canada), with all coefficients statistically different from -1, the Zipf coefficient (see footnote 26). The R-squared values are very high, with a minimum of 0.97 (Mexico), indicating a very good fit for each group. We also formally test whether each group's distribution follows a power law using the method from Gabaix and Ibragimov (2011), which tests the linearity of the log rank log population relationship by adding a quadratic term. Seven of the ten groups pass this test for the samples used in Table 5, while the distributions for India and for Canada pass the test when restricted to the top 65% and top 55% of neighborhoods, respectively. The rank-size plot for Mexican neighborhoods is somewhat curved and we reject that the distribution follows a power law unless we restrict to neighborhoods larger than the 85th percentile. Thus, with the exception of Mexican neighborhoods, we find strong evidence that the distribution of ethnic neighborhoods follows a power law for the ten largest groups in the 2010 census.

Power laws describe many economic phenomena and Zipf's law is perhaps the most well-known empirical regularity in the distribution of cities (Rozenfeld, Rybski, Gabaix, and Makse 2011, Gabaix 2016). While we do not find a coefficient of negative one for any of the ten groups, it's reassuring that our measure of neighborhoods follows a power law while a simpler tract-based measure does not. Mechanically, any measure of ethnic neighborhoods based on single tracts without clustering is unlikely to conform to a power law because tract populations are capped by the census. More meaningfully, many ethnic tracts are spatially

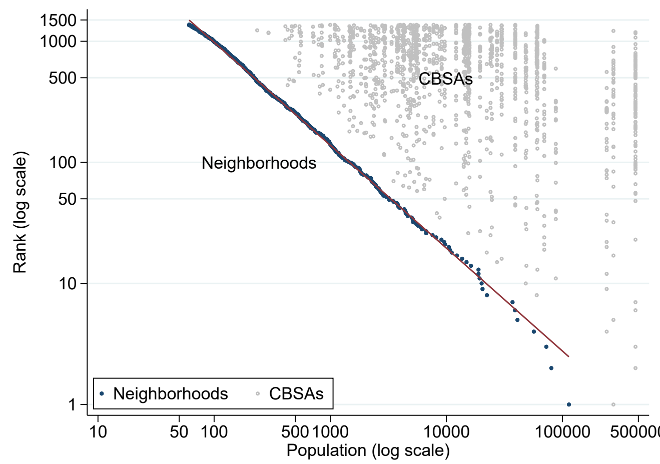
[27]Regressing log neighborhood rank on the log of the CBSA ethnic population alone leads to an R-squared of less than 0.15 for the ten largest groups we studied. If we add this term to the regression of log neighborhood rank on log neighborhood population, then the coefficient on neighborhood population changes by less than 0.01.

[28]It is not uncommon that power law relationships start above a minimum threshold. For example, Rozenfeld, Rybski, Gabaix, and Makse (2011) studies the distribution in US population clusters and find that a power law begins above the 25th percentile. Note that given the neighborhood distribution, excluding the bottom quartile of neighborhoods removes only a very small portion of the total population living in neighborhoods. For example, the bottom quartile of Chinese neighborhoods only contains 1.6% of the Chinese population living in Chinese neighborhoods.

Figure 4: Neighborhood Population vs Rank: Chinese Neighborhoods 2010



(a) Neighborhoods and Tracts



(b) Neighborhoods and Corresponding CBSAs

Panel a: The circular markers show rank of Chinese neighborhoods (rank 1 is largest neighborhood population) graphed against neighborhood Chinese population. The triangular markers (curve with higher intercept) plot the rank of Chinese tracts against the Chinese population in the tract. For each curve we plot the best fit line; the neighborhood fit line has a slope of -0.78 and an R-squared of 0.981. Both axes are scaled in natural logarithms. **Panel b:** The darker circular markers again show Chinese neighborhood rank plotted against neighborhood population, but now restricted to neighborhoods above the 25th percentile of Chinese population. The corresponding fit line has a slope of -0.85 and an R-squared of 0.998. The lighter circular markers to the right of the line show the Chinese population of the CBSA corresponding to each neighborhood, i.e., we plot the rank of Chinese neighborhood r against the total Chinese population of the containing CBSA.

Table 4: Neighborhood Size over Time

Year	Mean Population	Median Population	Mean Tract Ct.	Median Tract Ct.
1970	18,310	1,858	37	13
1980	60,211	4,265	68	19
1990	59,847	4,895	63	20
2000	62,300	9,399	61	27
2010	42,425	4,032	44	12

Notes: Columns 1 and 3 show the neighborhood population (ethnic population) and the count of tracts for the average foreign-born resident living in an ethnic neighborhood. Columns 2 and 4 show the same statistics but for the *median* foreign-born resident.

Table 5: Neighborhoods: Log Rank Regressed on Log Population

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Mexico	China	Philip	India	Vietna	ElSalv	Korea	Cuba	Canada	DominR
log pop	-0.70***	-0.86***	-0.95***	-0.92***	-0.97***	-0.93***	-0.99***	-1.07***	-1.39***	-0.79***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	1466	1378	1771	1427	1477	1363	1445	1312	1818	925
R^2	0.971	0.996	0.989	0.991	0.996	0.989	0.994	0.987	0.983	0.994

Notes: The dependent variable is the log of neighborhood rank, minus 1/2; see equation 10. For each group we restrict the sample to neighborhoods larger than the 25th percentile. Using the test from (Gabaix and Ibragimov 2011), we fail to reject that the distribution of ethnic neighborhoods follows a power law for every group, except for Mexican, Indian, and Canadian neighborhoods (i.e., 7 of 10 groups pass our test for a power law). Standard errors in parentheses with significance levels: *** 1 percent, ** 5 percent, * 10 percent.

clustered. If economic interactions between residents of these tracts lead to a power law, then the poor fit with a power law suggests tract-based measures fail to capture true neighborhood sizes.[29].

3.4.1 Large Neighborhoods

News outlets occasionally profile growing immigrant neighborhoods or lament the decline of historical ones.[30] However, it's difficult to know whether these patterns in specific neighborhoods represent general trends in the United States. In Appendix section 7.3 we try to answer this question by providing counts of large neighborhoods—those with at least 1000 people from an ethnic group—over time for many immigrant groups. Consistent with media reports, we find a decrease in the number of large neighborhoods for immigrants from Europe, especially Italy, Germany, Hungary, Ireland, and the former Yugoslavia. On the other hand, the count and size of neighborhoods for more recent Asian immigrants have (including Chinese, Indian, Korean, Filipino, and Vietnamese) have grown quickly.

[29]Dingel, Miscio, and Davis (2021) show that the distribution of administratively-defined cities in India and China do not follow a power law, but cities constructed using nighttime lights—a measure of economic activity—do.

[30]Typical examples include "Miami-Dade's Cubans: Where They Live, What They Think" (New York Times, December 21st, 2014), "Where Have All America's Little Italys Gone?" (NBC News, October 13, 2014), and "Vietnamese Forged a Community in New Orleans. Now It May Be Fading." (New York Times, May 5th, 2018).

4 Ethnic Neighborhood Spatial and Temporal Structure

Canonical urban economics theory suggests that if there is some benefit to the most central area of a region—the highest concentration of jobs or commerce—then the population will monotonically decrease from this point as a result of residents seeking to reduce travel time to the center (Fujita, Krugman, and Venables 1999). Ethnic neighborhoods may also function like small coherent regions—city sub-centers—with some type of advantage in the center, such as ethnic employment, ethnic retail, or even just the highest density of co-ethnic interaction. If this is the case then we should observe ethnic population gradients within a neighborhood, sloping downward from the densest point.[31] Additionally, increases in neighborhood population size should expand the geographic footprint of the neighborhood, incorporating new areas further from the center. To evaluate these hypotheses, we estimate neighborhood population gradients and examine when tracts in a neighborhood first attained ethnic tract status ("join year"), as a function of distance to the neighborhood center. We then examine how two characteristics of the housing stock, the rental percentage and age, vary with the ethnic tract year to look for patterns in the spatial expansion of the neighborhoods.

In Figure 5 we first illustrate our analysis with two plots of the largest Vietnamese neighborhood in 2010, located in Orange County, CA (117 contiguous tracts and 96,269 Vietnam-born residents). In panel A we plot the Vietnamese population of each neighborhood tract against the distance from the center of the neighborhood. We define the center of an ethnic neighborhood as the weighted sum of its component tract centroids (latitude, longitude), using each tract's share of the total neighborhood ethnic population as weights. The plot shows a strong negative population gradient where tracts further from the center have lower Vietnamese populations. In panel B we plot the first year each tract in the 2010 neighborhood was classified as a Vietnamese tract, or "join year" for joining the neighborhood. This plot shows that locations classified as ethnic tracts earlier tend to be closer to the center, although the pattern is not as strong as that in panel A.

Turning to the estimation, we regress an outcome $y_{jt,b}$ for tract j belonging to neighborhood b in year t on the logarithm of the distance between the centroid of j and the centroid of its neighborhood, b:

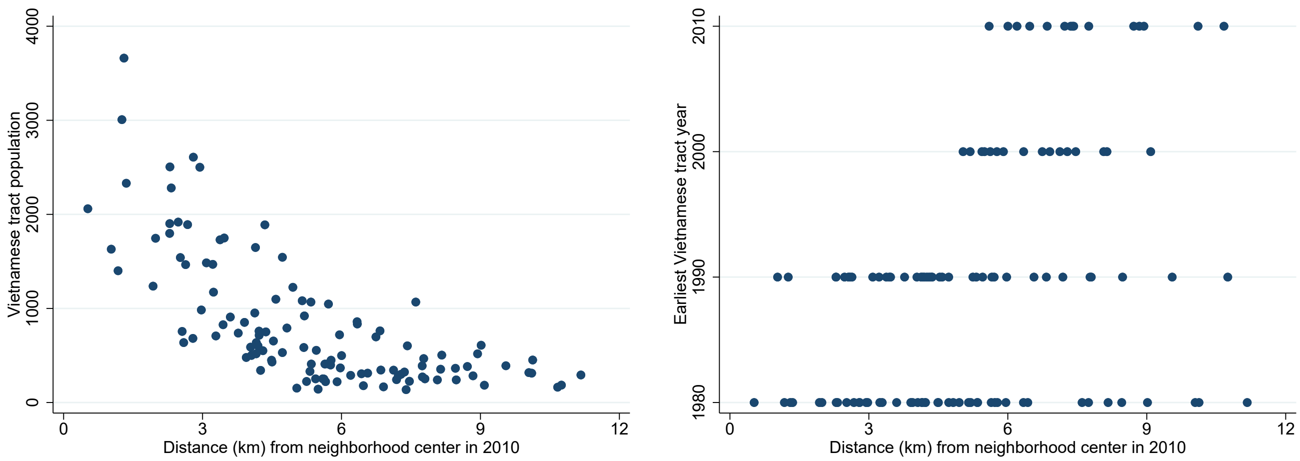
$$y_{jt,b} = \beta * ln(dist)_{jt,b} + \mu_b + \varepsilon_{jt,b} \tag{11}$$

Neighborhoods vary greatly in average population, age, and distance between component tracts, and so we include a neighborhood fixed effect, μ_b, in the specification. Note that a neighborhood b is defined as a set of contiguous tracts $j \in b$ for one group g in a specific year t; we do not try to define a unique neighborhood over multiple years. We use the logarithm of distance to the center, rather than distance in levels, so that differences in geographic tract size across neighborhoods do not affect the interpretation of our estimates.

An issue with using the above specification to estimate population gradients is that by defining the neighborhood centroid using each neighborhood tract's ethnic population, we may induce a downward sloping population gradient by construction. Even if tract populations are randomly distributed within each neighborhood, the population-weighted center is likely to be closer to tracts with larger populations. Therefore, to assess how much of the observed gradient could result from this mechanical effect alone, we use the following permutation procedure. Within each neighborhood, we randomly assign the set of observed tract ethnic populations to tracts. As a data operation, this is just a permutation (random sort) of tract populations while fixing the tract identifier within a neighborhood, thus the total neighborhood population is the same, but the population of each tract may differ from the actual data. We then calculate the (population-weighted) neighborhood centroid from the permuted data and run specification 11 to get an estimate of β that would result from randomly distributing tract populations within neighborhoods. We repeat this procedure 1000

[31] See (McMillen 2001) for a discussion of employment sub-centers and empirical identification.

Figure 5: Large Vietnamese Neighborhood in Orange County, CA, 2010



(a) Population

(b) Vietnamese Tract Join Year

Notes: Panel A plots the Vietnamese population in each tract against the tract's distance to the center of the 2010 neighborhood. Panel B plots the earliest year each tract in the 2010 neighborhood was classified as a Vietnamese tract ("join year"). Note that 1980 is the first year we have data on Vietnamese populations.

times and then report the 5th smallest and 995th smallest β estimates as the lower bound and upper bound of the 99% confidence interval under the null hypothesis of random locations.

In panel A of Table 6 we show the results for six groups and report the bounds of the 99% confidence interval from the permutation exercise in the last two table rows. The dependent variable is the logarithm of tract ethnic population and therefore the coefficient β is an elasticity reflecting the percentage change in population resulting from a percentage change in distance from the center. We restrict the sample to neighborhoods with at least five tracts and cluster standard errors at the neighborhood level.[32] For each group we find significant negative gradients that are below—and mostly far below—the lower bound of the permutation confidence interval. For example, a one percent increase in the distance to the center of a Vietnamese neighborhood is associated with a 0.29% decrease in the Vietnamese population. Many groups have elasticities above 0.2, while Canadian neighborhoods have the smallest elasticity at 0.1.

In panel B the dependent variable is the first year a 2010 neighborhood tract was classified as an ethnic tract ("join year"). The sample is restricted to 2010 neighborhoods, and only those 2010 neighborhoods with at least five component tracts. The coefficient is a semi-elasticity with the interpretation that a 1% increase in distance from the center results in an (approximately) $\beta/100$ increase in years. For example, if one tract is ten percent further from the center than another tract in a Chinese neighborhood, then we predict the further tract became a Chinese tract 0.282 years later. To better understand the magnitude of these coefficients we use specification 11 to predict the join year of the tract closest to the center of each neighborhood and the tract furthest from the center, and then average the difference across all neighborhoods (one observation per neighborhood). Thus, for the average Chinese neighborhood, we predict the closest tract became a Chinese tract about five and a half years earlier than the furthest tract. This is a modest but consistent effect across all groups, and leads to larger differences for geographically bigger neighborhoods, which also tend to be

[32]We chose five as the minimum neighborhood size that would yield at least 100 unique permutations within a neighborhood ($5! = 120$). The pattern of estimates is not that sensitive to this choice, with higher thresholds leading to slightly larger gradient estimates.

older. Lastly, note that the magnitude of a group's coefficient depends in part upon the earliest year in which the census tracked the group; we can track 2010 Italian neighborhoods back to 1970 but 2010 Vietnamese neighborhoods only to 1980. Therefore, we do not emphasize differences between groups, but rather that for all groups we find evidence that tracts further from the center of a neighborhood became ethnic tracts later.

In Table 3 we found that ethnic tracts tend to have more rental housing and an older housing stock than other locations where the ethnic group lives. We next examine how these characteristics change within a neighborhood by running a version of specification 11 in which we replace the log distance variable with the ethnic tract year—the dependent variable in panel B of Table 6—and use the rental percentage and percentage of the housing stock older than 30 years as outcomes. As in panel B, we constrain the sample to 2010 neighborhoods with at least five component tracts. In panel C we find that for every group, tracts that became part of the neighborhood later have younger housing. Similarly, in panel D we find that these tracts also have a lower percentage of rental housing. In Table 3 we also found that rental percentage and housing age were decreasing for most groups over time. This pattern could result from an extensive margin effect where newer neighborhoods have different characteristics, but the results in panels C and D show that these changes in the housing stock are also present within neighborhoods.

Summarizing the results in Table 6, panels A and B support the notion that ethnic neighborhoods have a structure like city sub-centers with spatial expansion as additional residents move to the neighborhood. However, tracts joining the neighborhood later have less rental housing and a newer housing stock. These changes could indicate that subsequent arrivals to a neighborhood are different (cohort effects) or that the neighborhood itself affects housing supply and maintenance. Nonetheless, the existence of such a pattern suggests that the surrounding housing stock could be an important factor in explaining how neighborhoods expand and that the central point of a neighborhood may offer some clues to the reasons for a specific neighborhood's formation.

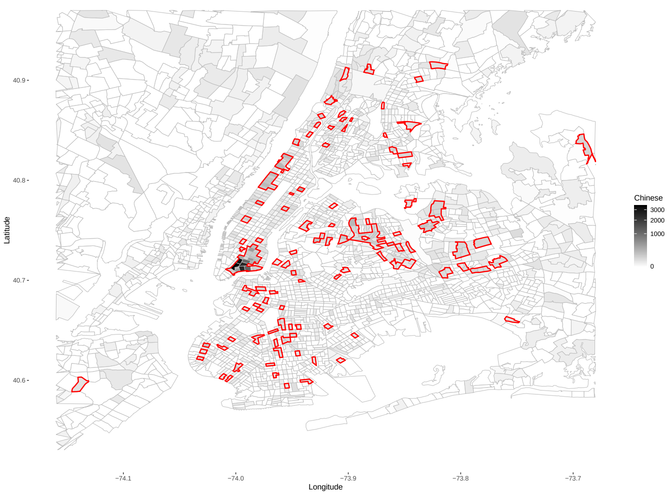
5 Dynamics of Ethnic Neighborhoods

In Figure 6 we show the same map of New York City Chinese neighborhoods as in panel B of Figure 2, but using data from 1970-2000. The (square root) scale is relative to the year, with dark black indicating a bit over 3000 people in 1970 but over 6000 people in 2000. The tracts composing Manhattan's Chinatown (roughly -74.00, 40.72, see panel A of Figure 2) have only changed slightly over time and the overall geographic size is roughly the same across all years. In fact, the population weighted center of the neighborhood moved just 280 meters from 1970 to 2010. On the other hand, the neighborhoods in Queens—Elmhurst ($-73.88, 40.74$) and Flushing ($-73.83, 40.75$)—have become much larger. Further, the large cluster of Chinese tracks in 2000, comprised of the Sunset Park ($-74.01, 40.64$) and Bensonhurst ($-74.00, 40.60$) neighborhoods, was just a few isolated Chinese tracts in 1970. Lastly, through the five decades there are many singleton tracts that appear and then disappear, never turning into neighborhoods of contiguous tracts. These patterns raise several questions. How long does an ethnic neighborhood last, how do ethnic neighborhoods grow, and how does the existing population of a neighborhood affect growth?

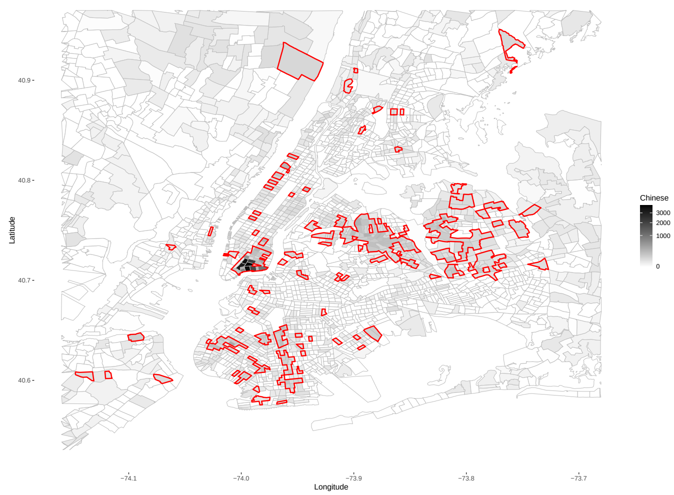
5.1 Neighborhood Attrition

We start to answer these questions by showing a simple transition matrix in Table 8 for the likelihood an ethnic tract in period t remains an ethnic tract in some future period. We do this in two ways, which we refer to as unweighted and weighted. Our unweighted measure is simply the share of all ethnic tracts from period t that are also ethnic tracts p years later: $tractcount_{t,t+p}/tractcount_t$. A tract remains an ethnic tract in $t+p$ if it is an ethnic tract for the same group in both t and $t+p$. For example, a Chinese tract in year

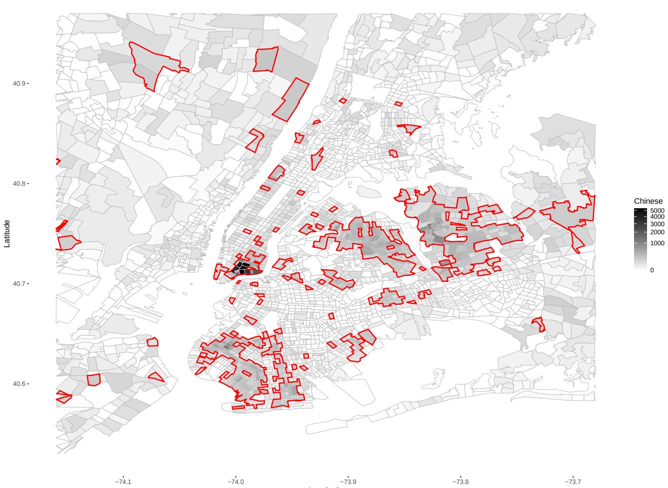
Figure 6: New York City Chinese Neighborhoods, 1970-2000



(a) 1970



(b) 1980



(c) 1990

(d) 2000

Notes: Each panel plots the same subset of NYC Census tracts, shading each tract by count of Chinese, using a square root scale specific to that year. The red borders show clusters of contiguous Chinese tracts defined using our algorithm.

Table 6: Ethnic Population, Ethnic Tract Year, and Housing within Neighborhoods

	(1) foreign	(2) Canada	(3) China	(4) Cuba	(5) Domin. R.	(6) India	(7) Italy	(8) Jamaica	(9) Mexico	(10) Vietnam
log distance	-0.12***	-0.10***	-0.24***	-0.32***	-0.28***	-0.17***	-0.22***	-0.29***	-0.25***	-0.29***
	(0.01)	(0.01)	(0.02)	(0.04)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	32992	3316	9069	4318	3235	8234	6237	4351	16430	6606
Clusters	1455	383	618	254	194	651	452	238	898	515
CI_lb	-0.04	-0.08	-0.09	-0.10	-0.10	-0.10	-0.07	-0.09	-0.08	-0.11
CI_ub	-0.01	-0.02	-0.04	-0.01	-0.01	-0.04	-0.02	-0.01	-0.03	-0.04

(a) Log Ethnic Population by Distance

	(1) foreign	(2) Canada	(3) China	(4) Cuba	(5) Domin. R.	(6) India	(7) Italy	(8) Jamaica	(9) Mexico	(10) Vietnam
log distance	3.67***	1.42**	2.82***	6.19***	4.01***	2.35***	4.92***	4.72***	4.37***	3.77***
	(0.37)	(0.66)	(0.41)	(1.01)	(0.48)	(0.33)	(0.73)	(0.55)	(0.58)	(0.42)
Observations	7256	1003	2989	777	1042	3016	822	1240	5394	2099
Clusters	351	120	196	57	60	223	70	71	302	157
Pred. Yr. Range	7.86	2.42	5.52	12.09	9.15	4.58	9.26	9.37	9.39	7.39

(b) Ethnic Tract Year by Distance

	(1) foreign	(2) Canada	(3) China	(4) Cuba	(5) Domin. R.	(6) India	(7) Italy	(8) Jamaica	(9) Mexico	(10) Vietnam
join year	-0.51***	-0.34***	-0.23***	-0.61***	-0.12**	-0.31***	-0.31***	-0.42***	-0.27***	-0.55***
	(0.05)	(0.08)	(0.05)	(0.18)	(0.05)	(0.06)	(0.05)	(0.09)	(0.04)	(0.11)
Observations	7193	1000	2971	776	1025	3009	820	1221	5364	2098
Clusters	351	120	196	57	60	223	70	71	302	157
D. Var. Mean	68.3	47.6	56.6	66.4	79.8	43.0	77.0	78.4	72.8	60.4

(c) Percentage of Housing Stock Older than 30 Years by Ethnic Tract Year

	(1) foreign	(2) Canada	(3) China	(4) Cuba	(5) Domin. R.	(6) India	(7) Italy	(8) Jamaica	(9) Mexico	(10) Vietnam
join year	-0.46***	-0.01	-0.27***	-0.45***	-0.51***	-0.30***	-0.36***	-0.28***	-0.44***	-0.44***
	(0.04)	(0.05)	(0.05)	(0.12)	(0.06)	(0.06)	(0.06)	(0.09)	(0.04)	(0.07)
Observations	7251	1003	2986	777	1042	3016	822	1240	5392	2099
Clusters	351	120	196	57	60	223	70	71	302	157
D. Var. Mean	50.9	35.1	41.3	49.9	66.9	35.8	38.0	49.0	52.8	41.5

(d) Rental Percentage by Ethnic Tract Year

Note: The dependent variable in panel A is log ethnic population. The sample covers all years but is restricted to neighborhoods with at least five tracts in that year. The bottom two rows of this panel report the lower bound and upper bound of a 99% confidence interval generated under the null hypothesis of random assignment of tract populations within a neighborhood. In panels B, C, and D the sample is restricted to neighborhoods in 2010 with at least five tracts. In panel B the dependent variable is the earliest year a tract attained ethnic tract status ("join year"). The last row of this panel shows the predicted age difference between the closest tract and further tract of a neighborhood, averaged across neighborhoods. In panel C the dependent variable is the percentage of the housing stock in a tract built more than 30 years earlier and in panel D the dependent variable is the percentage of a tract's stock that is rental housing. Both variables are measured in percentage points, zero to 100. The independent variable in panels C and D is the join year—the dependent variable in panel B. All specifications in both panels include neighborhood fixed effects and standard errors are clustered at the neighborhood level, see equation 11 in the text.

t must also be a Chinese tract in $t+p$.[33] Our weighted measure seeks to answer the question: for a given

[33] We restrict the sample to birthplace countries tracked by the census in both t and $t+p$. Since the number of birth countries tracked has expanded over time, this restriction excludes very few groups. Across all $(t, t+p)$ pairs shown in Table 8 the average

ethnic resident of an ethnic tract in period t, what is the likelihood the tract they reside in is also an ethnic tract in $t+p$? To do so, we weight each tract in period t by its share of the total population living in ethnic tracts in period t. In Table 8 the years in the leftmost column show the initial year while the cells in the table show the share remaining ethnic tracts in subsequent years. The first row of the table shows that only 37% of ethnic tracts in 1970 remained ethnic tracts in 1980, and only 18% of 1970 tracts were also ethnic tracts in 2010. The unweighted share of tracts remaining ethnic tracts in the following decade is fairly consistent across years, averaging 36%. The drop-off in subsequent periods is far less than in the first decade and again, is consistent over time. The second row for each year shows the weighted share and is much higher than the unweighted share. This suggests that ethnic tracts with larger (ethnic) populations are more likely to remain ethnic tracts in later periods. However, even this measure drops quickly with a 30% decrease in the first period and further decreases of about nine percentage points in each subsequent period. These statistics show that ethnic neighborhoods change quickly and the majority of ethnic tracts lose their status after one decade. Famous historical neighborhoods, like New York City's Manhattan Chinatown, are rather exceptional and most small ethnic neighborhoods can be expected to disappear after a few decades.

Table 8: Ethnic Tract Transition Matrix

Initial Year	Type	1980	1990	2000	2010
1970	unweighted	0.37	0.27	0.23	0.18
	weighted	0.66	0.54	0.46	0.38
1980	unweighted		0.39	0.33	0.27
	weighted		0.72	0.64	0.57
1990	unweighted			0.35	0.28
	weighted			0.71	0.62
2000	unweighted				0.32
	weighted				0.71

Notes: The cells in the "unweighted" rows show the share of ethnic tracts in the initial year that remain ethnic tracts in each subsequent decade. The cells in the "weighted" rows weight each tract by its share of the total population living in ethnic tracts in the initial year. For a resident of an ethnic tract in the intial year, this measure is equal to the likelihood that the tract they reside in is still an ethnic tract in a subsequent decade.

5.2 Neighborhood Growth

Next, we look at the growth of ethnic neighborhoods by examining the change in the group population for tracts that are part of a neighborhood and tracts that border neighborhoods. In Table 9 we show the results based on the specification below, run separately for a selection of groups from different regions of the world and for whom we have at least four decades of census data.

$$\triangle_{(t+1)} n_{cj} = \beta_1 * ethtract_{cjt} + \beta_2 * adj.ethtract_{cjt} + \beta_3 * nbpop_{c,b(jt)} + \beta_4 * adj.nbpop_{c,a(jt)} + \mu_{ct} + \varepsilon_{cjt} \quad (12)$$

The dependent variable is the change in the group population of a census tract j in city c, over two consecutive periods $\triangle_t$ (one decade): $n_{cj,t+1} - n_{cj,t}$. We use the change in levels, rather than logs, because there are many tract-years where a group population is zero. The variable $ethtract_{cjt}$ indicates whether the tract is an ethnic tract in the current period while $adj.ethtract_{cjt}$ indicates whether the tract is adjacent to an

coverage is 98% of the ethnic tracts in year t.

ethnic tract, but not an ethnic tract itself, in the current period. We use the notation $b(j,t)$ to represent the neighborhood b of tract j in year t and $a(j,t)$ to represent the neighborhood immediately adjacent (shares a border) to tract j. The variable $nbpop_{c,b(jt)}$ gives the neighborhood population if j is in a neighborhood in t; this variable is zero if j is not an ethnic tract in the current period. The variable $adj.nbpop_{c,a(jt)}$ captures the population of the adjacent neighborhood, if j borders a neighborhood and is not an ethnic tract in the current period. Again, this variable is zero if j does not border a neighborhood. Both population variables are measured in thousands of people. We include CBSA-by-year fixed effects and therefore the coefficients represent values relative to other tracts in the same city and year. For each group we first run a specification estimating β_1 and β_2 only, which shows the average growth in ethnic tracts and tracts bordering ethnic tracts. We then estimate all four coefficients to explore whether the ethnic tract's neighborhood population, or the population of a bordering neighborhood, have an additional effect on growth. In both specifications we cluster standard errors at the census tract level.

Column 1 of the top panel of Table 9 shows that, on average, native tracts lose 270 natives in the following decade and that locations bordering native tracts lose 111 natives, in comparison to tracts that are not native tracts and do not border native tracts. The coefficients in the second column show that native tracts in larger neighborhoods, and tracts bordering larger neighborhoods, lose more natives in the following period. Thus, native tracts, and adjacent tracts, are losing population relative to other tracts in the city. The patterns are similar for Italian neighborhoods and Canadian neighborhoods, although the latter are modestly expanding into adjacent tracts with an additional 1.3 people, relative to other tracts. On the other hand, Chinese, Indian, and Mexican neighborhoods are growing rapidly. Chinese tracts gain an additional 27 Chinese residents and tracts adjacent to Chinese tracts gain 12 residents. Further, tracts in large Chinese neighborhoods, and adjacent to large neighborhoods, grow faster. Each additional 1000 people in a Chinese neighborhood raises the per-tract growth by 2.3 people and the growth in all adjacent (non-Chinese) tracts by 0.5 people. Similar patterns hold for Indian neighborhoods. Mexican neighborhoods are also growing, but the -0.23 coefficient on the neighborhood population implies that the largest neighborhoods are growing more slowly. Controlling for population, Vietnamese and Jamaican ethnic tracts lose residents, relative to other tracts in the same city-year. However, these neighborhoods may still be growing through expansion into neighboring tracts, which would be consistent with the overall growth shown for these groups in Appendix Table 12.

Some readers may wonder how much of the variation in growth in Table 9 could be explained with tract population, or the populations of adjacent tracts. In short, conditional on simple population measures, do ethnic tract and neighborhood designations provide any additional explanatory power for understanding growth? In Appendix Table 14 we first estimate the effect of ethnic tract status on growth, controlling for both the ethnic population of the tract and the sum of the adjacent tracts' ethnic populations. For nearly all groups, ethnic tract status is still an important predictor of growth. Not surprisingly, the coefficient on ethnic tract status is much smaller than in Table 9 since ethnic population mostly determines ethnic tract status. We then estimate the full specification in equation 12 but add the ethnic population of the tract as a control, and again find that neighborhood populations and ethnic tract status predict growth. Ethnic tract status is determined by ethnic population and total population—used as a proxy for housing capacity—and thus some of the additional explanatory power may come from controlling for this capacity. Additionally, as noted in section 3.2, clusters of ethnic tracts are not perfectly circular, and therefore neighborhood population size may do a better job capturing spatial effects than simply summing the adjacent tracts' population.

Like the results in Table 6, a general pattern shown by Table 9 is that many neighborhoods grow by expanding geographically into neighboring tracts. This is consistent with the work in Saiz and Wachter (2011) who show that census tracts bordering places with many immigrants (from all ethnicities) receive more immigrants in the following decade. But how much of the growth occurs through expansion (extensive margin) versus growth within tracts already part of the neighborhood (intensive margin)?

As mentioned earlier, we do not try to define unique neighborhoods and follow them over time; our definition of a neighborhood is only valid in a given period. Instead, we take a neighborhood defined in period t and partition the total population change into tracts that were already ethnic tracts in the previous period $t-1$ (intensive margin) and those that became ethnic tracts in the current period t (extensive margin). Let I_b denote the set of tracts in neighborhood b that were ethnic tracts in the previous period and E_b be the tracts that were not. The ethnic population change of the neighborhood is the sum of the population change in each neighborhood tract: $\triangle_t n_b = \sum_{k \in I_b} \triangle_t n_k + \sum_{k \in E_b} \triangle_t n_k$. The extensive margin share of neighborhood population change is then the sum from new ethnic tracts divided by the total, or $\sum_{k \in E_b} \triangle_t n_k / \triangle_t n_b$. We show the average extensive margin share across neighborhoods and over time for eight groups in Table 10, restricting the sample to neighborhoods with at least five tracts and those with positive population change, $\triangle_t n_b > 0$. The mean is calculated over all neighborhoods—each neighborhood is a single observation—and we show the standard error in parentheses. For all groups, most of the population increase of a neighborhood comes from new ethnic tracts. In fact, for several groups the share in the extensive margin is larger than one because for these groups intensive margin tracts are actually losing population on average.[34] These results show that neighborhoods grow into adjacent tracts and the majority of population growth occurs through spatial expansion.

6 Conclusion

In this paper we derived a new statistical definition of an ethnic neighborhood from a choice model and using the native population as a reference distribution. We then applied our definition to census tract data in the United States from 1970-2010, across many immigrant groups. We found that a large and increasing percentage of the foreign-born population reside in ethnic neighborhoods, from 43% in 1970 to 67% in 2010. These neighborhoods are systematically different from other locations where the ethnic population lives, with lower household incomes and housing values, as well as a higher percentage of rental housing and residents commuting without a car. Most neighborhoods last only a couple decades with 70% of ethnic tracts disappearing after a single decade, but large neighborhoods can persist longer. The number of large neighborhoods for many European groups has declined precipitously. However, for many Asian and Latin American groups, ethnic neighborhoods are growing rapidly. For groups with large immigration flows, existing neighborhood locations predict the settlement locations of new arrivals and larger neighborhoods attract larger inflows. Neighborhood growth occurs primarily through spatial expansion into adjacent locations. Additionally, we find that ethnic neighborhoods have striking similarities to empirical patterns documented in cities. Many neighborhoods have a spatial structure similar to a city sub-center, with a population density gradient declining from a central point with the oldest tracts in the neighborhood. Further, the neighborhood population distribution follows a power law with the biggest neighborhoods orders of magnitude larger than the smallest neighborhoods.

We used summary-level (census tract) data for our analysis, but our results raise a number of additional and intriguing questions that might be answerable with individual-level data. Some of the patterns we have found in ethnic neighborhoods, including the high percentage of people commuting without a car, the percentage of renters, and the age of the housing stock, suggest that transportation and housing policy could affect whether and where neighborhoods form. In fact, a recent article in The Economist magazine suggests that the Danish government is using public housing policy to try and discourage the growth of ethnic

[34] The extensive margin share and intensive margin share must sum to one, thus the intensive margin shares are negative on average for Canada, Cuba, and Italy (-0.03,-0.92,-3.37).

Table 9: Ethnic Neighborhoods and Growth

	(1) Native	(2) Native	(3) Canada	(4) Canada	(5) China	(6) China	(7) India	(8) India
ethtract	-269.64***	-101.18***	-29.23***	-26.91***	27.30***	11.43***	15.57***	5.41***
	(24.46)	(36.49)	(0.44)	(0.46)	(1.77)	(1.32)	(1.44)	(1.22)
adjacent to ethtract	-110.65***	-66.81***	1.34***	1.36***	11.77***	10.46***	11.74***	10.12***
	(12.04)	(16.46)	(0.12)	(0.14)	(0.30)	(0.29)	(0.31)	(0.29)
neighborhood pop		-42.78***		-5.86***		2.25***		4.84***
		(9.02)		(1.20)		(0.16)		(0.53)
adjacent neighborhood pop		-12.33***		-0.10		0.52***		1.65***
		(3.56)		(0.44)		(0.06)		(0.17)
Observations	244268	244268	244292	244292	244292	244292	192238	192238
Neigh. Pop Mean		3.94		0.40		7.29		2.18
Adj. Neigh. Pop Mean		3.54		0.21		1.94		0.94

(a) Selected Groups

	(1) Italy	(2) Italy	(3) Jamaica	(4) Jamaica	(5) Mexico	(6) Mexico	(7) Vietnam	(8) Vietnam
ethtract	-43.27***	-38.35***	0.65	-3.75***	112.18***	124.92***	-1.20	-10.50***
	(0.73)	(0.69)	(1.20)	(1.06)	(2.65)	(3.01)	(1.45)	(1.10)
adjacent to ethtract	-0.80***	-0.77***	6.37***	5.20***	45.66***	41.91***	9.79***	8.34***
	(0.13)	(0.13)	(0.28)	(0.24)	(1.01)	(0.98)	(0.31)	(0.29)
neighborhood pop		-1.44***		0.71***		-0.23***		2.71***
		(0.14)		(0.13)		(0.02)		(0.25)
adjacent neighborhood pop		-0.08		0.84***		0.21***		1.09***
		(0.06)		(0.10)		(0.02)		(0.13)
Observations	244292	244292	192238	192238	244292	244292	192238	192238
Neigh. Pop Mean		3.60		6.84		56.11		3.56
Adj. Neigh. Pop Mean		1.17		1.41		17.62		1.16

(b) Selected Groups

Note: Dependent variable in each regression is future difference in ethnic population, $ethpop_{t+1} - ethpop_t$. All regressions include CBSA-by-year fixed effects. The variable "ethtract" indicates the status of the tract in the current period while "adjacent to ethtract" indicates the tract is next to an ethnic tract (but not one itself). The "neighborhood pop" is the population of the neighborhood in the current period for an ethnic tract (zero for non-ethnic tracts). The adjacent neighborhood population measures the population for non-ethnic tracts that border a neighborhood in the current period (zero for tracts not bordering neighborhoods). Both population variables are measured in thousands; the last two rows of each table show the means over all non-zero cases. Standard errors are clustered at the tract level.

Table 10: Share of Neighborhood Growth in New Ethnic Tracts

	(1) Canada	(2) China	(3) Cuba	(4) India	(5) Italy	(6) Jamaica	(7) Mexico	(8) Vietnam
Ext. Share	1.03	0.80	1.92	0.74	4.37	0.88	0.61	0.87
	(0.04)	(0.10)	(0.44)	(0.05)	(2.65)	(0.14)	(0.03)	(0.05)
N	297	559	159	568	152	184	801	433

Note: We partition a neighborhood's tracts in decade t into those that were already ethnic tracts in the previous decade (intensive margin) and those that became an ethnic tract in t (extensive margin). We sum the ethnic population change between periods among all neighborhood tracts and then calculate the share of the neighborhood change in extensive margin tracts ("Ext. Share"). We use all years of data but limit the sample to neighborhoods with at least five tracts and an increase in total population. We then calculate the mean over all neighborhoods—each neighborhood is one observation—and put the standard error in parentheses. For some groups, intensive margin tracts lose population on average, and thus the extensive share is larger than one (see text).

neighborhoods.[35] This leads to a fundamental question: how do ethic neighborhoods start? There is likely tremendous heterogeneity in neighborhood formation, but detailed characteristics on the first generation to settle in a location might reveal some common patterns. Relatedly, how long does the first generation reside in these neighborhoods and where does the second generation locate? Finally, most studies on the causal effect of ethnic neighborhoods use larger geographic definitions or coarser measures of neighborhoods, and thus it could be worthwhile to re-examine some of these outcomes using a more precise neighborhood definition. Answering any of the questions above requires detailed data on neighborhood residents, often only available in confidential datasets. Nonetheless, for researchers with access to this type of data, these could be promising areas for future work.

[35] "Denmark wants to break up ethnic enclaves. What is wrong with them?", *The Economist*, November 28, 2019

References

AARONSON, D., D. A. HARTLEY, AND B. MAZUMDER (2017): "The effects of the 1930s HOLC Redlining maps," .

ABRAMITZKY, R., L. P. BOUSTAN, AND D. CONNOR (2020): "Leaving the Enclave: Historical Evidence on Immigrant Mobility from the Industrial Removal Office," *NBER Working Paper 27372*.

AGARWAL, S., H.-S. CHOI, J. HE, AND T. F. SING (2019): "Matching in housing markets: The role of ethnic social networks," *The Review of Financial Studies*, 32(10), 3958–4004.

ALBA, R. D., J. R. LOGAN, B. J. STULTS, G. MARZAN, AND W. ZHANG (1999): "Immigrant groups in the suburbs: A reexamination of suburbanization and spatial assimilation," *American sociological review*, pp. 446–460.

ALLEN, J. P., AND E. TURNER (1996): "Spatial patterns of immigrant assimilation," *The Professional Geographer*, 48(2), 140–155.

ANANAT, E. O. (2011): "The wrong side (s) of the tracks: The causal effects of racial segregation on urban poverty and inequality," *American Economic Journal: Applied Economics*, 3(2), 34–66.

ANSELIN, L. (1995): "Local indicators of spatial association—LISA," *Geographical analysis*, 27(2), 93–115.

ASLUND, O., P.-A. EDIN, P. FREDRIKSSON, AND H. GRONQVIST (2011): "Peers, Neighborhoods, and Immigrant Student Achievement: Evidence from a Placement Policy," *American Economic Journal: Applied Economics*.

BAUER, T., G. S. EPSTEIN, AND I. N. GANG (2005): "Enclaves, language, and the location choice of migrants," *Journal of Population Economics*.

BAYER, P., F. FERREIRA, AND R. MCMILLAN (2007): "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 115(4), 588–638.

BENNETT, J. (2020): "The share of immigrant workers in high-skill jobs is rising in the U.S.," Discussion paper, Pew Research Center.

BERTRAND, M., E. F. LUTTMER, AND S. MULLAINATHAN (2000): "Network Effects and Welfare Cultures*," *The Quarterly Journal of Economics*, 115(3), 1019–1055.

BILLINGS, S. B., AND E. B. JOHNSON (2012a): "The location quotient as an estimator of industrial concentration," *Regional Science and Urban Economics*, 42(4), 642–647.

BILLINGS, S. B., AND E. B. JOHNSON (2012b): "A non-parametric test for industrial specialization," *Journal of Urban Economics*, 71(3), 312–331.

BLUMENBERG, E., AND A. E. EVANS (2010): "Planning for demographic diversity: the case of immigrants and public transit," *Journal of Public Transportation*, 13(2), 2.

BORJAS, G. (1995): "Ethnicity, Neighborhoods, and Human-Capital Externalities," *The American Economic Review*, pp. 365–390.

BOSCH, M., M. A. CARNERO, AND L. FARRE (2010): "Information and discrimination in the rental housing market: Evidence from a field experiment," *Regional science and urban Economics*, 40(1), 11–19.

BUDIMAN, A., C. TAMIR, L. MORA, AND L. NOE-BUSTAMANTE (2020): "Facts on U.S. Immigrants, 2018," Discussion paper, Pew Research Center.

CHATMAN, D. G., AND N. KLEIN (2009): "Immigrants and travel demand in the United States: implications for transportation policy and future research," *Public Works Management & Policy*, 13(4), 312–327.

CHETTY, R., AND N. HENDREN (2018a): "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects," *The Quarterly Journal of Economics*, 133(3), 1107–1162.

——— (2018b): "The impacts of neighborhoods on intergenerational mobility II: County-level estimates," *The Quarterly Journal of Economics*, 133(3), 1163–1228.

CHETTY, R., N. HENDREN, AND L. F. KATZ (2016): "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment," *American Economic Review*.

CHISWICK, B. R., AND P. W. MILLER (1996): "Ethnic networks and language proficiency among immigrants," *Journal of Population Economics*, 9(1), 19–35.

CUTLER, D., E. GLAESER, AND J. VIGDOR (2008a): "Is the melting pot still hot? Explaining the resurgence of immigrant segregation," *The Review of Economics and Statistics*, 90(3), 478–497.

——— (2008b): "When are ghettos bad? Lessons from immigrant segregation in the United States," *Journal of Urban Economics*, 63(3), 759–774.

CUTLER, D. M., AND E. L. GLAESER (1997): "Are ghettos good or bad?," *The Quarterly Journal of Economics*, 112(3), 827–872.

DAMM, A. P. (2009): "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence," *Journal of Labor Economics*, 27(2), 281–314.

——— (2014): "Neighborhood quality and labor market outcomes: Evidence from quasi-random neighborhood assignment of immigrants," *Journal of Urban Economics*.

DE BELLEFON, M.-P., P.-P. COMBES, G. DURANTON, L. GOBILLON, AND C. GORIN (2020): "Delineating urban areas using building density," *Journal of Urban Economics*, p. 103226.

DINGEL, J. I., A. MISCIO, AND D. R. DAVIS (2021): "Cities, lights, and skills in developing economies," *Journal of Urban Economics*.

DURANTON, G., AND H. OVERMAN (2005): "Testing for localization using micro-geographic data," *Review of Economic Studies*, pp. 1077–1106.

EDIN, P., P. FREDRIKSSON, AND O. ÅSLUND (2003): "Ethnic enclaves and the economic success of immigrants—Evidence from a natural experiment," *The Quarterly Journal of Economics*, 118(1), 329–357.

ELLISON, G., AND E. GLAESER (1997): "Geographic Concentration in US Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, 105(5).

FEE, K. D., AND D. A. HARTLEY (2013): "The Relationship between City Center Density and Urban Growth or Decline," in *Revitalizing American Cities*, ed. by S. Wachter, and K. Zeuli, chap. 3. University of Pennsylvania Press.

FUJITA, M., P. KRUGMAN, AND A. J. VENABLES (1999): *The Spatial Economy*. MIT Press.

GABAIX, X. (2016): "Power Laws in Economics: An Introduction," *Journal of Economic Perspectives*.

GABAIX, X., AND R. IBRAGIMOV (2011): "Rank - 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents," *Journal of Business and Economics Statistics*.

GLAESER, E. L., AND J. GYOURKO (2005): "Urban Decline and Durable Housing," *Journal of Political Economy*, 113(2).

GOLD, S. J. (2015): "Ethnic Enclaves," *Emerging Trends in the Social and Behavioral Sciences*.

GUIMARÃES, P., O. FIGUEIREDO, AND D. WOODWARD (2007): "Measuring the localization of economic activity: a parametric approach," *Journal of Regional Science*, 47(4), 753–774.

GUIMARÃES, P., O. FIGUEIREDO, AND D. WOODWARD (2009): "Dartboard tests for the location quotient," *Regional Science and Urban Economics*, 39(3), 360–364.

HELLERSTEIN, J. K., M. J. KUTZBACH, AND D. NEUMARK (2014): "Do labor market networks have an important spatial dimension?," *Journal of Urban Economics*, 79, 39–58.

HIRSCHMAN, C., AND D. S. MASSEY (2008): "Places and Peoples: The New American Mosaic," in *New Faces in New Places: The Changing Geography of American Immigration*, ed. by D. S. Massey, chap. 1. Russel Sage Foundation.

IOANNIDES, Y. M., AND L. D. LOURY (2004): "Job Information Networks, Neighborhood Effects, and Inequality," *Journal of Economic Literature*.

KEEGAN, T. H., E. M. JOHN, K. M. FISH, T. ALFARO-VELCAMP, C. A. CLARKE, AND S. L. GOMEZ (2010): "Breast cancer incidence patterns among California Hispanic women: differences by nativity and residence in an enclave," *Cancer Epidemiology and Prevention Biomarkers*, 19(5), 1208–1218.

LAZEAR, E. P. (1999): "Culture and language," *Journal of political Economy*, 107(S6), S95–S126.

LI, W. (1998): "Anatomy of a New Ethnic Settlement: The Chinese Ethnoburb in Los Angeles," *Urban studies*, 35(3), 479–501.

LIM, S., S. Y. STELLA, N. L. DE LA CRUZ, AND C. TRINH-SHEVRIN (2017): "Defining ethnic enclave and its associations with self-reported health outcomes among Asian American adults in New York City," *Journal of immigrant and minority health*, 19(1), 138–146.

LOGAN, J., W. ZHANG, AND R. ALBA (2002): "Immigrant enclaves and ethnic communities in New York and Los Angeles," *American Sociological Review*, pp. 299–322.

LOGAN, J. R. (2013): "The Persistence of Segregation in the 21st Century Metropolis," *City & Community*, 12(2).

LOGAN, J. R., J. XU, AND B. STULTS (2012): "Interpolating US Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database," *Professional Geographer*.

LOUF, R., AND M. BARTHELEMY (2016): "Patterns of Residential Segregation," *PLOS One*.

MANSON, S., J. SCHROEDER, D. V. RIPER, AND S. RUGGLES (2018): "IPUMS National Historical Geographic Information System: Version 13.0 Database," .

MCMILLEN, D. P. (2001): "Nonparametric employment subcenter identification," *Journal of Urban Economics*, 50(3), 448–473.

MORENO-MONROY, A. I., M. SCHIAVINA, AND P. VENERI (2020): "Metropolitan areas in the world. Delineation and population trends," *Journal of Urban Economics*, p. 103242.

MORI, T., K. NISHIKIMI, AND T. E. SMITH (2005): "A divergence statistic for industrial localization," *Review of Economics and Statistics*, 87(4), 635–651.

MORI, T., AND T. E. SMITH (2013): "A probabilistic modeling approach to the detection of industrial agglomerations," *Journal of Economic Geography*, 14(3), 547–588.

MUNSHI, K. (2003): "Networks in the modern economy: Mexican migrants in the US labor market," *The Quarterly Journal of Economics*, 118(2), 549–599.

OSYPUK, T. L., A. V. D. ROUX, C. HADLEY, AND N. R. KANDULA (2009): "Are immigrant enclaves healthy places to live? The Multi-ethnic Study of Atherosclerosis," *Social science & medicine*, 69(1), 110–120.

PORTES, A., AND L. JENSEN (1989): "The Enclave and the Entrants: Patterns of Ethnic Enterprise in Miami before and after Mariel," *American Sociological Review*.

PORTES, A., AND R. G. RUMBAUT (2014): *Immigrant America, 4th edition*. University of California Press.

ROZENFELD, H. D., D. RYBSKI, X. GABAIX, AND H. A. MAKSE (2011): "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*.

SAIZ, A. (2010): "The geographic determinants of housing supply," *The Quarterly Journal of Economics*, 125(3), 1253–1296.

SAIZ, A., AND S. WACHTER (2011): "Immigration and the Neighborhood," *American Economic Journal: Economic Policy*, 3(2), 169–88.

SANDERS, J. M., AND V. NEE (1992): "Problems in Resolving the Enclave Economy Debate," *American Sociological Review*.

SCHIFF, N. (2015): "Cities and product variety: evidence from restaurants," *Journal of Economic Geography*.

VOGEL, K. B., R. GOLDBLATT, G. H. HANSON, AND A. K. KHANDELWAL (2020): "Detecting Urban Markets with Satellite Imagery: An Application to India," *Journal of Urban Economics*.

WALDINGER, R. (1993): "The Ethnic Enclave Debate Revisited," *International Journal of Urban and Regional Research*.

WONG, M. (2013): "Estimating Ethnic Preferences Using Ethnic Housing Quotas in Singapore," *Review of Economic Studies*.

ZHOU, M., Y.-F. TSENG, AND R. Y. KIM (2008): "Rethinking residential assimilation: The case of a Chinese ethnoburb in the San Gabriel Valley, California," *Amerasia Journal*, 34(3), 55–83.

7 Appendix

7.1 Calculating Adjusted P-values

In theory one could use many draws from the null distribution (natives) to calculate p-values for every tract, and then another set of draws to generate a set of family-level p-values. However, in practice we found this procedure too slow to generate an adjusted p-value for each ethnic group in each city in each year. Instead, we use a statistical software implementation of the multinomial distribution from the base functions of the R programming language.

To find the adjusted p-value for ethnic tracts we first draw N_g people from a multinomial distribution with category probabilities equal to the native shares, $s_{0,1},...,s_{0,J}$, where $s_{0,j} = n_{0,j}/N_0$. This gives a vector of simulated tract counts for group g, $n_{g1},...,n_{gJ}$. Let i denote the index of the 99th percentile share, $\bar{s}$. For each of the tracts from 1 to i, we calculate the tract p-value as $1 - Pr(X \leq n_{gj})$ using the binomial CDF with N_g draws and success probability $\bar{s}$. Define the minimum p-value from these i tracts as $pmin_r$, where r indicates replication; the remaining $J-i$ tracts are unused. We then repeat this procedure $R = 50,000$ times to get a distribution of family-level p-values, $pmin_1,...,pmin_R$. The adjusted p-value for specialization is the 5th percentile of these R values.

7.2 English Language Ability

In section 3.2 we showed the association between speaking the origin country language and living in an ethnic neighborhood. In this section we follow up by examining English language ability, measured as the fraction of those language speakers who report speaking English "less than very well" in the 2010 census. We use the following specification:

$$lowEngFrac_{cj} = \beta_1 * spkEthLang_{cj} + \beta_2 * ethtract_{cj} + \beta_3 * (spkEthLang_{cj} \times ethtract_{cj}) + \mu_c + \varepsilon_{jc} \quad (13)$$

The Census asks those who report speaking a particular (non-English) language at home whether they speak English "very well" or "less than very well," and reports the answers as counts. We take the count speaking "less than very well" and divide by the count speaking the ethnic language at home, $spkEthLang_{cj}$, and thus the dependent variable is always between 0 and 1 (we restrict the sample to locations where $spkEthLang_{cj} \geq 1$). The number of other speakers in the tract could affect language use, and so we also control for $spkEthLang_{cj}$. Our hypothesis is that β_2 will be positive, even conditional on the count of fellow speakers, since limited English ability could be one reason for choosing to live in an ethnic tract. We also include the interaction term and CBSA fixed effects. As before, we also test an additional specification where we drop the ethnic tract terms and include the adjacent ethnic population and the containing neighborhood size.

In Table 11 we find that the fraction of ethnic language speakers reporting low English ability is significantly higher in ethnic tracts for every group. In Chinese ethnic tracts, the fraction is 5.7 percentage points higher while the count of language speakers has no significant effect, in ethnic tracts or other locations. The "nb-var mean" row shows the mean count of language speakers in ethnic tracts in the odd columns (ex: 320 in Chinese tracts), and the mean size of the containing neighborhood for the even columns (ex: 11,050 for Chinese neighborhoods). For Italians and Koreans (columns 3 and 5), we do find significant effects on the count of speakers, but these coefficients roughly counterbalance each other (negative outside of ethnic tracts, positive and of the same magnitude in ethnic tracts), so that the main effect is still on the simple ethnic tract indicator. Turning to the neighborhood size specifications (even columns), we find that the containing neighborhood size is significantly associated with English ability, even conditional on immediately adjacent populations, for all groups except Vietnamese. Overall, these results are broadly consistent with the idea that ethnic tract status captures a non-linear effect of ethnic population on spoken language and English ability, and similarly, neighborhood population size captures spatial effects beyond simply summing immediately adjacent areas.

Table 11: Fraction of Language Speakers with Low Level English

	(1) China	(2) China	(3) Italy	(4) Italy	(5) Korea	(6) Korea	(7) Vietnam	(8) Vietnam
language speakers 000's	0.026	0.043**	-0.578***	-0.031	-0.408***	-0.057***	-0.068	0.020
	(0.033)	(0.018)	(0.153)	(0.040)	(0.116)	(0.016)	(0.120)	(0.021)
ethnic tract	0.057***		0.090***		0.023**		0.081***	
	(0.006)		(0.014)		(0.009)		(0.008)	
l. speakers X eth. tract	0.013		0.505***		0.411***		0.081	
	(0.038)		(0.173)		(0.113)		(0.118)	
adj. eth. pop. 000's		0.001		0.024		0.013**		0.023*
		(0.005)		(0.023)		(0.005)		(0.012)
neigh. pop. 000's		0.001***		0.028***		0.003***		0.000
		(0.000)		(0.002)		(0.001)		(0.000)
Observations	17866	17866	8564	8564	11850	11850	11144	11144
Clusters	130	130	45	45	105	105	103	103
dep-var mean	0.50	0.50	0.32	0.32	0.56	0.56	0.58	0.58
nb-var mean	0.32	11.05	0.11	0.95	0.18	5.42	0.20	6.93
Adj. R2	0.02	0.01	0.03	0.01	0.01	0.01	0.02	0.01

Notes: Dependent variable is the fraction of people speaking the language at home, who report speaking English "less than very well." The variable "language speakers" is the count speaking the language in the tract, measured in thousands. The table row nb-var mean reports the mean count of language speakers (000's) in ethnic tracts for odd columns; for even columns this is the mean of the containing neighborhood population. The sample is 2010 census tracts with at least one member of the ethnic group and at least one speaker of the language, in cities with at least 1000 people in the ethnic group. All specifications include CBSA fixed effects and standard errors are clustered by CBSA.

7.3 Large Neighborhood Profiles

In Table 12 we present three statistics for many immigrant groups over the five decades of our panel. One challenge in this exercise is choosing the appropriate unit of analysis. For example, if immigrant group A has five group tracts in a city, four clustered into one neighborhood and a singleton, what is the average neighborhood tract size? We think it's most informative to present the neighborhood size *for an average ethnic tract*, or the average weighted by tract count. In the example above, the weighted tract count is 17/5—4/5 of tracts are in a neighborhood of size 4 while 1/5 of tracts are in neighborhood of size 1—rather than the 5/2 calculated using the neighborhood as the unit of analysis. However, we also list the count of unique neighborhoods with a population of 1000 group residents or more. The precise count of neighborhoods in any year is sensitive to the process used to reconcile census tract boundaries, as well as edge issues, but the trends over time are still informative. We restrict the sample to city-years where the group population was at least 1000 people.

The first row of each group cell in Table 12 shows the average neighborhood population and the second row gives the average number of tracts in the neighborhood, both weighted by tract count. The third row gives the count of unique neighborhoods larger than 1000 group residents. For example, the average Chinese tract in 1970 was in a neighborhood that had 8.5 total tracts and a population of 1621 Chinese-born residents. By 2010 the average Chinese tract was in a neighborhood of 10,808 Chinese residents with 22.2 tracts. In 1970 there were only 12 unique neighborhoods with more than 1000 Chinese but in 2010 there were 142. Since native ethnic tracts are generally not spatially clustered, the count of tracts in a native neighborhood is very low. The population is high simply because each tract has thousands of natives. On the other hand, using "foreign" as a grouping variable, we find evidence that internationals from different ethnic groups tend to cluster. Some of the effect likely comes from the clustering of specific ethnic groups—a Chinatown is also a foreign cluster—but the count of tracts and neighborhood population counts are larger than could result from any single ethnic group. Consistent with media reports, the number of Italian neighborhoods with more than 1000 Italians has declined precipitously from 78 in 1970 to 24 in 2010. Moreover, the average Italian tract was in a neighborhood with 6926 Italians and 25.9 tracts in 1970 but only 880 Italians and 8.1 tracts in 2010. This pattern is similar for other European groups, including immigrants from Germany, Hungary, Ireland, and (the former) Yugoslavia. In contrast, the count and size of neighborhoods for more recent Asian immigrants has grown quickly, such as those for Chinese, Indian, Korean, Filipino, and Vietnamese. Immigrant neighborhoods for people from the Caribbean (Cuba, Dominican Republic, Haiti, Jamaica, Trinidad) tend to be quite large. While the count of Cuban neighborhoods above 1000 Cubans is fairly small, the average Cuban tract is in a neighborhood with 16,864 Cubans and 28.5 tracts in 1970, and 36,645 Cubans in 21.6 tracts in 2010. This results from very large Cuban neighborhoods in the Miami and New York City CBSAs. Lastly, for Mexicans, the largest immigrant group, the average neighborhood appears to be decreasing in size while the count of neighborhoods increases. This suggest Mexican immigrants may be moving away from the largest neighborhoods while still concentrating in smaller ones.

Table 12: Ethnic Neighborhoods (Clustered Tracts) Over Time

Birth Country	Variable	1970	1980	1990	2000	2010
Native	Nbhd. Pop.	4031	4280	3697	3770	3966
	Nbhd. Tract Ct.	1.5	1.7	1.5	1.4	1.1
	Nbhd.'s Abv. 1000	325	363	417	442	495
Foreign	Nbhd. Pop.	54956	107939	112471	115091	71559
	Nbhd. Tract Ct.	65.4	96.4	82.4	69.4	43.3
	Nbhd.'s Abv. 1000	380	404	521	698	1040
Canada	Nbhd. Pop.	542	381	321	516	372
	Nbhd. Tract Ct.	4	4	4.3	7.5	4.5
	Nbhd.'s Abv. 1000	16	15	8	18	20
China	Nbhd. Pop.	1621	3099	8175	12113	10808
	Nbhd. Tract Ct.	8.5	19.5	29	31.2	22.2
	Nbhd.'s Abv. 1000	12	18	55	99	142
Colombia	Nbhd. Pop.	.	.	3963	3863	2639
	Nbhd. Tract Ct.	.	.	22	17.1	11.8
	Nbhd.'s Abv. 1000	.	.	19	31	40
Cuba	Nbhd. Pop.	16864	34210	45532	51415	36645
	Nbhd. Tract Ct.	28.5	37.7	38.9	33.7	21.6
	Nbhd.'s Abv. 1000	20	21	19	20	26
Dominican Republic	Nbhd. Pop.	.	5409	11841	22582	18432
	Nbhd. Tract Ct.	.	15.1	30.6	40.8	30.3
	Nbhd.'s Abv. 1000	.	19	26	43	52
Egypt	Nbhd. Pop.	.	.	173	331	362
	Nbhd. Tract Ct.	.	.	3.7	5	4.1
	Nbhd.'s Abv. 1000	.	.	2	4	6
Ethiopia	Nbhd. Pop.	.	.	237	535	863
	Nbhd. Tract Ct.	.	.	5.1	6.8	6.8
	Nbhd.'s Abv. 1000	.	.	1	6	13
Germany	Nbhd. Pop.	2986	1048	387	243	159
	Nbhd. Tract Ct.	21.8	10.5	5.4	4	2.8
	Nbhd.'s Abv. 1000	17	18	14	7	4
Greece	Nbhd. Pop.	1205	1696	1277	1319	722
	Nbhd. Tract Ct.	10.4	13.3	13.3	14.2	7.8
	Nbhd.'s Abv. 1000	10	12	10	9	6
Haiti	Nbhd. Pop.	.	.	10726	14199	11708
	Nbhd. Tract Ct.	.	.	40.3	44.9	34.5
	Nbhd.'s Abv. 1000	.	.	21	33	41
Hungary	Nbhd. Pop.	807	576	274	214	95
	Nbhd. Tract Ct.	12.8	10.6	5.7	4.9	2.3
	Nbhd.'s Abv. 1000	13	6	5	2	1
India	Nbhd. Pop.	.	472	1426	4049	6145
	Nbhd. Tract Ct.	.	7.7	13.6	22	18.1
	Nbhd.'s Abv. 1000	.	7	29	82	161
Iran	Nbhd. Pop.	.	.	4286	4597	4189
	Nbhd. Tract Ct.	.	.	28.5	23	17.2
	Nbhd.'s Abv. 1000	.	.	10	15	17
Ireland	Nbhd. Pop.	2087	1292	1036	757	326
	Nbhd. Tract Ct.	16	14.3	12.2	10.1	4.7
	Nbhd.'s Abv. 1000	19	12	9	8	6
Italy	Nbhd. Pop.	6926	4811	2740	1718	880
	Nbhd. Tract Ct.	25.9	23.8	16.9	13.4	8.1
	Nbhd.'s Abv. 1000	78	66	41	29	24
Jamaica	Nbhd. Pop.	.	6049	7815	12855	10357
	Nbhd. Tract Ct.	.	31.1	34.6	43.8	33.8
	Nbhd.'s Abv. 1000	.	17	29	32	34
Japan	Nbhd. Pop.	710	636	1423	1460	896
	Nbhd. Tract Ct.	9.1	8.3	13.2	13.5	7.6
	Nbhd.'s Abv. 1000	9	11	20	18	20
Korea	Nbhd. Pop.	.	1616	3409	4812	4970
	Nbhd. Tract Ct.	.	14.5	18.2	21.6	16.6
	Nbhd.'s Abv. 1000	.	10	40	50	69
Laos	Nbhd. Pop.	.	.	1151	1305	1024
	Nbhd. Tract Ct.	.	.	9.2	10.4	7.4
	Nbhd.'s Abv. 1000	.	.	23	15	14
Mexico	Nbhd. Pop.	21742	68878	71814	60953	49055
	Nbhd. Tract Ct.	54.7	83.2	68.3	50.7	41
	Nbhd.'s Abv. 1000	42	110	186	368	536
Nigeria	Nbhd. Pop.	.	.	395	1061	911
	Nbhd. Tract Ct.	.	.	8.5	13.1	8.1
	Nbhd.'s Abv. 1000	.	.	3	7	17
Pakistan	Nbhd. Pop.	.	.	385	1074	806
	Nbhd. Tract Ct.	.	.	6.9	10.4	6.1
	Nbhd.'s Abv. 1000	.	.	6	16	26
Peru	Nbhd. Pop.	.	.	894	1353	885
	Nbhd. Tract Ct.	.	.	10.2	10.5	5.8
	Nbhd.'s Abv. 1000	.	.	9	21	27
Philipines	Nbhd. Pop.	.	4066	6263	5875	5070
	Nbhd. Tract Ct.	.	24.1	23.9	20.8	14.3
	Nbhd.'s Abv. 1000	.	27	50	73	95
Poland	Nbhd. Pop.	6224	4159	3671	6674	5774
	Nbhd. Tract Ct.	35.1	27.5	19.2	19.4	16.6
	Nbhd.'s Abv. 1000	34	31	25	25	21
Portugal	Nbhd. Pop.	1275	2882	4376	3972	2574
	Nbhd. Tract Ct.	6.9	11.5	16.8	15	9.7
	Nbhd.'s Abv. 1000	16	27	30	26	19
Russia	Nbhd. Pop.	.	.	.	4962	2377
	Nbhd. Tract Ct.	.	.	.	28.9	13.4
	Nbhd.'s Abv. 1000	.	.	.	27	22
Thailand	Nbhd. Pop.	.	.	297	457	316
	Nbhd. Tract Ct.	.	.	5.6	6	3.4
	Nbhd.'s Abv. 1000	.	.	5	11	8
Trinidad	Nbhd. Pop.	.	.	7125	22410	9772
	Nbhd. Tract Ct.	.	.	45.6	136.1	50.6
	Nbhd.'s Abv. 1000	.	.	6	5	7
United Kingdom	Nbhd. Pop.	448	325	443	451	409
	Nbhd. Tract Ct.	4.8	4.3	6.5	6.3	5.1
	Nbhd.'s Abv. 1000	13	6	10	10	10
Vietnam	Nbhd. Pop.	.	749	3833	6002	6514
	Nbhd. Tract Ct.	.	7.5	16.7	17.9	16
	Nbhd.'s Abv. 1000	.	14	42	81	92
Yugoslavia*	Nbhd. Pop.	912	1478	680	534	221
	Nbhd. Tract Ct.	10	15.6	9	8.6	3.5
	Nbhd.'s Abv. 1000	10	9	9	7	5

Notes: The first row of each group reports the average neighborhood population of an ethnic tract and the second row gives the average neighborhood tract count. The third row is the count of unique neighborhoods with more than 1000 people. For each group, the sample is restricted to cities that had at least 1000 people from the group in that year.
*Yugoslavia dissolved in 2003 but was still a response category for place of birth in the 2006-2010 American Community Surveys. We use this category in 2010 to be consistent with earlier years.

7.4 Additional Results

Table 13: Housing Characteristics of Ethnic Tracts

Birth Country	Statistic	1970	1980	1990	2000	2010	Birth Country	Statistic	1970	1980	1990	2000	2010
	Rental Pct.	-0.012	0.055	0.046	0.033	0.003		Rental Pct.	.	.	0.067	0.037	0.039
	se	0.017	0.021	0.022	0.024	0.022		se	.	.	0.032	0.021	0.02
	Cmt No Car Pct.	0.003	0.05	0.042	0.043	0.002		Cmt No Car Pct.	.	.	0.065	0.038	0.016
	se	0.008	0.012	0.013	0.019	0.017		se	.	.	0.005	0.006	0.004
	Med HH Inc.	-0.19	-3.461	-5.178	-7.204	-3.631		Med HH Inc.	.	.	-5.536	-6.861	-7.867
Native	se	0.254	0.655	1.546	2.205	2.563	**Haiti**	se	.	.	1.805	1.036	1.795
	Med. H. Rent	-0.072	-0.437	-0.523	-0.848	-0.535		Med. H. Rent	.	.	-0.499	-0.681	-0.694
	se	0.034	0.046	0.144	0.212	0.215		se	.	.	0.244	0.204	0.197
	Med. H. Value	-0.009	-0.133	-0.163	-0.241	-0.214		Med. H. Value	.	.	-0.261	-0.385	-0.313
	se	0.005	0.02	0.073	0.113	0.095		se	.	.	0.026	0.031	0.081
	H. Age Gt. 30 Yrs Pct.	0.022	0.121	0.068	0.104	0.126		H. Age Gt. 30 Yrs Pct.	.	.	0.14	0.129	0.101
	se	0.014	0.015	0.018	0.028	0.058		se	.	.	0.033	0.022	0.024
	Rental Pct.	0.221	0.213	0.201	0.201	0.163		Rental Pct.	.	0.082	0.053	0.062	0.033
	se	0.021	0.017	0.016	0.013	0.011		se	.	0.027	0.012	0.01	0.008
	Cmt No Car Pct.	0.14	0.125	0.099	0.074	0.047		Cmt No Car Pct.	.	0.057	0.023	-0.004	-0.009
	se	0.025	0.026	0.022	0.02	0.018		se	.	0.022	0.009	0.004	0.005
	Med HH Inc.	-2.073	-3.916	-6.775	-11.307	-14.352		Med HH Inc.	.	-0.911	-0.514	2.254	4.113
Foreign	se	0.219	0.551	1.383	1.464	1.661	**India**	se	.	0.645	0.935	0.84	0.9
	Med. H. Rent	-0.153	-0.3	-0.437	-0.59	-0.691		Med. H. Rent	.	-0.073	0.066	0.644	0.559
	se	0.026	0.071	0.15	0.153	0.146		se	.	0.051	0.093	0.121	0.08
	Med. H. Value	-0.015	-0.056	-0.062	-0.208	-0.332		Med. H. Value	.	-0.029	0.002	0.047	0.068
	se	0.003	0.023	0.054	0.044	0.056		se	.	0.016	0.037	0.048	0.06
	H. Age Gt. 30 Yrs Pct.	0.201	0.149	0.087	0.073	0.063		H. Age Gt. 30 Yrs Pct.	.	0.021	-0.031	-0.105	-0.093
	se	0.02	0.019	0.019	0.015	0.014		se	.	0.016	0.014	0.014	0.01
	Rental Pct.	0.071	0.022	0.014	-0.009	0.008		Rental Pct.	0.105	0.068	0.043	0.021	0.026
	se	0.011	0.011	0.012	0.014	0.013		se	0.015	0.018	0.015	0.012	0.008
	Cmt No Car Pct.	0.035	0.026	0.025	0.019	0.028		Cmt No Car Pct.	0.092	0.063	0.038	0.019	0.017
	se	0.01	0.011	0.016	0.019	0.016		se	0.02	0.02	0.016	0.01	0.005
	Med HH Inc.	0.237	1.988	4.53	11.054	7.27		Med HH Inc.	-1.931	-2.501	-4.171	-3.296	-3.855
Canada	se	0.245	0.545	0.813	1.243	1.274	**Italy**	se	0.123	0.498	0.952	1.414	1.017
	Med. H. Rent	0.068	0.194	0.464	1.299	0.686		Med. H. Rent	-0.186	-0.256	-0.248	-0.056	-0.017
	se	0.023	0.047	0.069	0.155	0.113		se	0.016	0.049	0.099	0.113	0.098
	Med. H. Value	0.021	0.149	0.391	0.738	0.573		Med. H. Value	-0.034	-0.061	0.019	0.098	0.162
	se	0.006	0.029	0.064	0.112	0.086		se	0.002	0.01	0.039	0.071	0.062
	H. Age Gt. 30 Yrs Pct.	0.056	0.024	0.016	-0.042	-0.016		H. Age Gt. 30 Yrs Pct.	0.194	0.16	0.109	0.063	0.038
	se	0.013	0.013	0.009	0.01	0.007		se	0.016	0.021	0.017	0.016	0.008
	Rental Pct.	0.168	0.115	0.083	0.079	0.051		Rental Pct.	.	0.056	0.057	0.043	0.041
	se	0.018	0.034	0.025	0.016	0.011		se	.	0.011	0.012	0.011	0.007
	Cmt No Car Pct.	0.156	0.13	0.095	0.063	0.048		Cmt No Car Pct.	.	0.103	0.083	0.06	0.039
	se	0.025	0.041	0.028	0.02	0.017		se	.	0.029	0.025	0.018	0.015
	Med HH Inc.	-2.187	-2.616	-2.786	0.133	0.543		Med HH Inc.	.	-3.322	-6.286	-8.062	-9.495
China	se	0.153	0.789	1.65	1.909	2.353	**Jamaica**	se	.	0.565	0.563	0.905	0.937
	Med. H. Rent	-0.183	-0.272	-0.186	0.391	0.177		Med. H. Rent	.	-0.25	-0.48	-0.532	-0.597
	se	0.023	0.057	0.127	0.165	0.125		se	.	0.03	0.057	0.11	0.111
	Med. H. Value	-0.013	-0.034	0.119	0.217	0.328		Med. H. Value	.	-0.145	-0.303	-0.372	-0.355
	se	0.004	0.021	0.031	0.048	0.048		se	.	0.012	0.038	0.042	0.06
	H. Age Gt. 30 Yrs Pct.	0.161	0.13	0.045	-0.026	-0.019		H. Age Gt. 30 Yrs Pct.	.	0.119	0.117	0.085	0.064
	se	0.024	0.024	0.016	0.014	0.01		se	.	0.017	0.014	0.016	0.018
	Rental Pct.	0.171	0.16	0.147	0.121	0.07		Rental Pct.	0.168	0.176	0.176	0.175	0.146
	se	0.021	0.021	0.021	0.02	0.013		se	0.016	0.016	0.011	0.007	0.005
	Cmt No Car Pct.	0.074	0.067	0.062	0.039	0.014		Cmt No Car Pct.	0.091	0.083	0.079	0.057	0.028
	se	0.019	0.016	0.011	0.007	0.005		se	0.013	0.011	0.011	0.012	0.008
	Med HH Inc.	-2.037	-3.739	-7.839	-9.815	-8.705		Med HH Inc.	-3.07	-6.326	-13.399	-17.09	-20.975
Cuba	se	0.195	0.359	0.941	1.327	1.495	**Mexico**	se	0.202	0.501	0.814	0.832	0.985
	Med. H. Rent	-0.184	-0.368	-0.766	-0.877	-0.58		Med. H. Rent	-0.377	-0.76	-1.255	-1.57	-1.684
	se	0.02	0.04	0.1	0.118	0.091		se	0.027	0.071	0.09	0.15	0.156
	Med. H. Value	-0.039	-0.127	-0.192	-0.288	-0.241		Med. H. Value	-0.061	-0.266	-0.49	-0.574	-0.788
	se	0.006	0.015	0.035	0.039	0.058		se	0.006	0.031	0.105	0.077	0.101
	H. Age Gt. 30 Yrs Pct.	0.155	0.157	0.124	0.115	0.068		H. Age Gt. 30 Yrs Pct.	0.212	0.232	0.184	0.16	0.142
	se	0.028	0.025	0.017	0.014	0.014		se	0.027	0.027	0.02	0.016	0.017
	Rental Pct.	.	0.165	0.194	0.191	0.166		Rental Pct.	.	0.13	0.097	0.062	0.031
	se	.	0.023	0.038	0.051	0.047		se	.	0.018	0.015	0.01	0.006
	Cmt No Car Pct.	.	0.136	0.134	0.121	0.092		Cmt No Car Pct.	.	0.079	0.055	0.029	0.006
	se	.	0.032	0.045	0.048	0.043		se	.	0.019	0.015	0.008	0.004
	Med HH Inc.	.	-5.241	-11.912	-16.004	-18.521		Med HH Inc.	.	-3.518	-6.364	-8.151	-7.997
Dominican R.	se	.	0.677	1.669	3.219	4.049	**Vietnam**	se	.	0.332	0.718	0.803	0.958
	Med. H. Rent	.	-0.427	-1.033	-1.345	-1.365		Med. H. Rent	.	-0.3	-0.408	-0.589	-0.48
	se	.	0.024	0.131	0.293	0.377		se	.	0.048	0.073	0.083	0.115
	Med. H. Value	.	-0.163	-0.37	-0.389	-0.463		Med. H. Value	.	-0.09	-0.169	-0.334	-0.342
	se	.	0.018	0.068	0.065	0.053		se	.	0.008	0.028	0.034	0.045
	H. Age Gt. 30 Yrs Pct.	.	0.169	0.076	0.068	0.072		H. Age Gt. 30 Yrs Pct.	.	0.119	0.086	0.062	0.028
	se	.	0.018	0.012	0.021	0.024		se	.	0.024	0.019	0.014	0.008

Table 14: Ethnic Neighborhoods and Growth, Controlling for Population

	(1) Native	(2) Native	(3) Canada	(4) Canada	(5) China	(6) China	(7) India	(8) India
ethtract	-347.15***	-350.55***	-4.33***	-4.66***	7.05**	5.46**	-10.56***	-6.61***
	(22.61)	(33.46)	(0.47)	(0.47)	(3.01)	(2.33)	(2.20)	(2.01)
ethnic pop	-252.09***	-244.24***	-567.81***	-514.78***	0.14	69.85***	103.38**	219.12***
	(2.96)	(2.54)	(9.86)	(8.01)	(36.10)	(24.70)	(41.60)	(36.53)
adjacent ethnic pop	3.52***		37.78***		34.45***		74.94***	
	(0.39)		(1.33)		(5.61)		(4.88)	
adjacent to ethtract		-94.07***		4.36***		10.00***		9.19***
		(15.61)		(0.14)		(0.32)		(0.33)
neighborhood pop		-0.43		5.98***		1.76***		2.55***
		(7.65)		(0.74)		(0.20)		(0.52)
adjacent neighborhood pop		2.58		2.80***		0.49***		1.46***
		(3.35)		(0.43)		(0.06)		(0.17)
Observations	244268	244268	244292	244292	244292	244292	192238	192238
Neigh. Pop Mean		3.94		0.40		7.29		2.18
Adj. Neigh. Pop Mean		3.54		0.21		1.94		0.94

(a) Selected Groups

	(1) Italy	(2) Italy	(3) Jamaica	(4) Jamaica	(5) Mexico	(6) Mexico	(7) Vietnam	(8) Vietnam
ethtract	-6.02***	-6.06***	-3.05	-2.12	104.39***	131.90***	-4.87***	-6.97***
	(1.03)	(1.05)	(1.90)	(1.74)	(3.83)	(3.95)	(1.72)	(1.76)
ethnic pop	-438.52***	-401.65***	-165.99***	-26.35	-89.09***	-20.29***	-226.56***	-49.05**
	(16.27)	(14.25)	(31.55)	(25.55)	(9.20)	(7.20)	(26.91)	(24.31)
adjacent ethnic pop	22.58***		46.83***		19.50***		70.39***	
	(1.82)		(3.84)		(1.46)		(4.28)	
adjacent to ethtract		2.21***		5.30***		42.38***		8.54***
		(0.16)		(0.26)		(1.00)		(0.30)
neighborhood pop		1.21***		0.87***		-0.19***		3.17***
		(0.14)		(0.17)		(0.03)		(0.29)
adjacent neighborhood pop		0.12*		0.85***		0.22***		1.13***
		(0.06)		(0.10)		(0.02)		(0.14)
Observations	244292	244292	192238	192238	244292	244292	192238	192238
Neigh. Pop Mean		3.60		6.84		56.11		3.56
Adj. Neigh. Pop Mean		1.17		1.41		17.62		1.16

(b) Selected Groups

Note: Dependent variable in each regression is future difference in ethnic population, $ethpop_{t+1} - ethpop_t$. All regressions include CBSA-by-year fixed effects. The variable "ethtract" indicates the status of the tract in the current period while "adjacent to ethtract" indicates tract is next to an ethnic tract (but not one itself). The "neighborhood pop" is the population of the neighborhood in the current period for an ethnic tract (zero for non-ethnic tracts). The adjacent neighborhood population measures the population for non-ethnic tracts that border a neighborhood in the current period (zero for tracts not bordering neighborhoods). Both population variables are measured in thousands; the last two rows of each table show the means over all non-zero cases. Standard errors are clustered at the tract level.