Gentrification and Pioneer Businesses

Kristian Behrens* Brahim Boualam† Julien Martin‡ Florian Mayneris§

January 11, 2022

Abstract

Little is known about where hotspots of gentrification emerge within a city, and the role that some types of businesses play in the process. We develop a method to detect the sectors whose presence heralds the process of gentrification in a neighborhood. We show that these sectors, mostly found in cultural and creative industries, help to anticipate neighborhood change and that their predictive power complements that of traditional gentrification determinants. We also examine mechanisms related to amenities, worker characteristics, and signaling, which are consistent with these results. The analysis illustrates the importance of businesses in the socio-demographic dynamics of neighborhoods.

Keywords: Gentrification; pioneer businesses; micro-geographic data; NY.

JEL Classification: R14, R23, R31.

*Corresponding author: ESG, Université du Québec à Montréal, Canada; HSE University, Russian Federation; and CEPR, UK. E-mail: behrens.kristian@uqam.ca

†Statistics Canada, Investment, Science and Technology Division. E-mail: brahim.boualam@statcan.gc.ca

‡ESG, Université du Québec à Montréal, Canada; and CEPR, UK. E-mail: martin.julien@uqam.ca

§ESG, Université du Québec à Montréal, Canada; and CEPR, UK. E-mail: mayneris.florian@uqam.ca

1 Introduction

The influx of affluent residents in poor neighborhoods and the skyrocketing housing prices that usually follow are major concerns in big cities around the world. Detecting early signs of gentrification is thus crucial to better anticipate these changes. This paper shows that the locations of specific businesses can help to better anticipate *where* hotspots of gentrification emerge within a city.

Popular perception and accounts of gentrification often involve stories featuring alternative cafés, jazz clubs, and art galleries. Although previous academic work has focused on demographic changes, amenities, and housing market dynamics in gentrifying areas, the literature is mostly silent on the composition of, and changes in, stores and businesses in these neighborhoods.[1] This paper is among the few to put businesses at the heart of the analysis of gentrification.[2] A novel aspect of our anal-

[1]Gentrification has been related to crime (O'Sullivan, 2005; Ellen et al., 2019), the life-cycle of the housing stock (Brueckner and Rosenthal, 2009), housing prices (Guerrieri et al., 2013), the displacement of incumbent residents (McKinnish et al., 2010; Ding et al., 2016), and educated workers' preferences (Edlund et al., 2015; Su, 2018). See Rosenthal and Ross (2015) for a survey.

[2]Some exceptions—usually qualitative case studies—exist among geographers and sociologists (see, e.g., Lees 2003, Zukin et al. 2009, or Sullivan and Shaw 2011 on retail gentrification; and Grodach et al. 2014 on commercial arts and gentrification). Few papers link gentrification to businesses in a quantitative framework. Lester and Hartley (2014) show that manufacturing jobs tend to be replaced by restaurant and retail service jobs in gentrifying neighborhoods. Schuetz (2014) and Meltzer (2016) analyze the dynamics of art galleries and small businesses in gentrifying neighborhoods, respectively. Couture and Handbury (2020) show that the young educated workers' taste for consumption amenities such as bars, restaurants, and gyms explains a significant share of the urban revival of American cities. Using Yelp data, Glaeser et al. (2018) investigate how changes in the number of grocery stores, cafés, restaurants, and bars help to explain the growth rate of housing prices and the share of educated residents at the ZIP-code level. These papers focus on *ad hoc* lists of sectors, whereas we estimate that list from the data.

ysis is the identification of economic sectors that are systematically associated with the process of gentrification in its early stages. Using data for New York in 1990, we find that 26—mostly cultural and creative—sectors are over-represented in neighborhoods that will gentrify between 1990 and 2000. We further show that the location of these sectors is a significant predictor of gentrification episodes, both in subsequent decades (New York between 2000 and 2010) and other cities (Philadelphia between 2000 and 2010). The focus on where gentrification occurs makes this paper complementary to recent analyses that shed light on the timing of, and macro-forces behind, the gentrification observed in many cities.[3]

The paper is organized in three parts. First, we present a method to detect *pioneer* sectors, defined as sectors that are usually found in affluent areas but tend to be overrepresented initially in poor areas that will gentrify over the next decade. We detect these pioneer sectors using geographically fine-grained data on businesses and residents and three alternative definitions of gentrification.

Second, we present the 26 pioneer sectors we detect. They mainly belong to the cultural and creative industries and echo anecdotal evidence from the popular press and case studies on gentrification. To our knowledge, this is the first list of sectors associated with gentrification built systematically from micro-data for a major city. We analyze the predictive power of these pioneers and show that their location helps to predict gentrification in New York and Philadelphia, even after controlling for a large set of neighborhood characteristics or adopting an instrumental variable

[3] Baum-Snow and Hartley (2016) and Couture and Handbury (2020) emphasize the role that the changing tastes of young educated workers play in the urban revival of U.S. cities. Couture et al. (2018) highlight the role of increased income inequality on the upgrading of downtown areas and analyze the welfare implications of gentrification. Brummet and Reed (2018) investigate whether changes associated with gentrification are driven by incumbents or by in-migration, which also matters for welfare.

strategy. Hence, the presence of these businesses contains relevant information on future neighborhood change that goes beyond the usual determinants identified in the literature.

Finally, we discuss several mechanisms that could explain why pioneer sectors herald gentrification. For example, some pioneer establishments might provide consumer amenities to future wealthy and educated residents and their presence may signal the upgrading of the neighborhood. We also find that their workers have the characteristics of gentrifiers—younger, more educated, less kids—and live nearer to their workplace, which may provide a link between local businesses and the traits of local residents. Finally, the presence of pioneer sectors could also attract high-income households who value the proximity to artists and creative people. Whereas these explanations imply a causal role of pioneers in the gentrification process, an alternative view is that their presence captures other unobserved drivers of gentrification, such as changes in preferences for some traits of poor neighborhoods. We show that the presence of pioneer sectors is a good proxy for those unobserved drivers, since they are not captured by the presence of other sectors with similar characteristics.

The remainder of the paper is organized as follows. Section 2 describes the strategy used to detect pioneer sectors. Section 3 presents the pioneer industries and shows our main econometric results. Section 4 discusses some mechanisms through which pioneers herald gentrification. Section 5 finally concludes. Detailed data descriptions and additional material are relegated to a supplemental online appendix.

2 Pioneer Sectors and Gentrification

In this section, we describe our strategy to detect sectors found in poor neighborhoods ahead of their gentrification. We also discuss how we measure gentrification.

Because of space constraints, we relegate a detailed discussion of the data in Appendices A and O.2.

2.1 Detecting Pioneer Sectors

Gentrification has been widely studied from the residents' perspective, but more rarely from the businesses' perspective (notable exceptions include Glaeser et al. 2018; Meltzer 2016; Meltzer and Ghorbani 2017). However, most popular accounts or case studies of gentrification mention businesses—such as upscale restaurants and cafés, art galleries, and jazz clubs—and how they accompany, or even herald, gentrification. Yet, there is little systematic evidence on which sectors interact with the gentrification process.

We fill this gap by proposing a methodology to detect sectors that are statistically associated with gentrification prior to its occurrence. More precisely, we investigate which sectors, usually present in high-income blocks, are initially over-represented in low-income blocks that will gentrify in the near future.

Formally, we run negative binomial regressions to assess whether the geographic distribution of establishments of some sector s in base year t is systematically related to gentrification episodes during Δt. The equation we estimate is:

$$n^s_{b,t} = F\Big(\alpha^{sg}_0 + \alpha^{sg}_1 \text{poor}^g_{b,t} + \alpha^{sg}_2 \text{gentri}^g_{b,t+\Delta t} + (X^g_{b,t})'\beta^{sg}\Big) + \nu^{sg}_{b,t} \tag{1}$$

where $n^s_{b,t}$ is the count of establishments in sector s located in block b in year t.[4] $\text{gentri}^g_{b,t+\Delta t}$ is a measure of gentrification over period Δt measured on neighbor-

[4]There is a high prevalence of zeroes in our data, i.e., blocks without any establishment from a given industry. Hence, using the establishment count rather than the share seems preferable. We we estimate a negative binomial rather than a Poisson model because the presence of zeros leads to substantial over-dispersion in the data.

hood g of block b while $\text{poor}_{b,t}^{g}$ is a dummy measuring the presence of poor blocks (defined here as blocks whose income per capita is below the median of the urban area) in year t in the neighborhood g of block b. Finally, $X_{b,t}^{g}$ are other neighborhood characteristics and $\nu_{b,t}^{sg}$ is a block-sector-neighborhood specific error term. Since census block boundaries change over time, we have constructed blocks with a stable geography across years.[5]

We estimate this model for $t = 1990$ using the National Establishment Time-Series (NETS) data for the New York core-based statistical area (CBSA). We group establishments by their 6-digit NAICS code and use the geographical information system (GIS) software to assign them to blocks using latitude and longitude information.

Our main explanatory variable is the measure of gentrification ($\text{gentri}_{b,t+\Delta t}^{g}$), which we discuss at length in Section 2.2. We control for other important determinants of establishments' location choices, namely operating costs and market potential. The former are proxied by the logarithm of residential rents in the block.[6] The latter is proxied using the logarithms of the initial population and employment in the neighborhood.

The location of a business is not entirely driven by the characteristics of its own block, but also by the influence of the surrounding areas. The right-hand side variables in equation (1) are thus measured at the neighborhood level g to account for the possible influence of nearby blocks on the dynamics of establishments within block b.[7] We define a block's neighborhood alternatively by: (i) contiguous blocks ($g = cont$); (ii) blocks with centroids less than 250 meters from the centroid of block

[5] Details on the construction of these blocks are given in the data appendix.

[6] Residential rents are an imperfect proxy for commercial rents, which we do not observe for our blocks. Yet, Kan et al. (2004) show there is a positive correlation between residential and commercial rents within blocks. We think that including the former as controls is better than no controls at all.

[7] Rents are the only exception. Since location decisions do not directly depend on prices in nearby blocks, we measure rents at the block level, i.e., at the same spatial scale as the dependent variable.

b ($g = 250m$); and (iii) blocks with centroids less than 500 meters from the centroid of block b ($g = 500m$).

The coefficients of interest in equation (1) are α_1^{sg} and α_2^{sg}. The former captures the propensity of establishments in sector s to locate in the surroundings of poor blocks. The latter captures the extent to which those establishments are overrepresented in the surroundings of poor blocks that gentrify over the next period.

In what follows, we focus on sectors for which $\widehat{\alpha}_1^s < 0$ and $\widehat{\alpha}_2^s > 0$ and where both are significant at the 1% level. These sectors are usually not found in the vicinity of poor blocks, except those that will gentrify over the next decade. In other words, these sectors seem to make 'atypical location choices' that carry informational content regarding future gentrification. We refer to establishments in those sectors as *pioneer businesses* or simply *pioneers*.[8]

2.2 Measuring Gentrification

Gentrification is commonly viewed as a process affecting poor neighborhoods that experience a substantial influx of wealthier and more educated residents over a given period (e.g., Freeman and Braconi, 2004; Zukin et al., 2009; McKinnish et al., 2010; Lester and Hartley, 2014). Yet, there is no consensus on how to measure it. This *"lack of consensus concerning the conceptualization of gentrification allowed researchers to identify gentrified neighborhoods in a variety of ways"* (Barton, 2016, p.3). In search of an operational definition of gentrification, we reviewed 27 academic papers in the online Appendix O.4. Our bibliometric analysis shows the most prominent dimensions

[8] Another reason we focus on sectors that do not usually locate in poor neighborhoods ($\widehat{\alpha}_1^s < 0$) is that we examine how the arrival of pioneers in a poor neighborhood predicts its future gentrification in Section 3.2. Sectors with $\widehat{\alpha}_1^s > 0$ would thus not be very informative. Still, we show in Section 3.1 that the list of sectors remains very similar if we relax this restriction.

used to define gentrification are changes in income and education. Most of the papers further retain an eligibility criterion: a neighborhood gentrifies if, starting from an *initially low* income level, it experiences a *substantial increase* in income and/or in the share of highly educated residents. Formally, we consider here that only poor blocks (defined as those with an initial per capita income below the median of the metropolitan area) are eligible.

We thus build three distinct measures of gentrification. The first measure $M1$ takes value 1 if a poor block b gentrifies over the period Δt, and 0 otherwise. In this definition, a block is considered gentrifying if: (i) it is initially poor; (ii) it moves at least three deciles upwards in the metropolitan per capita income distribution during Δt; and (iii) it moves at least one decile upwards in the metropolitan distribution of the share of educated residents (those with a college degree) over the same period.[9]

Measure $M1$ is multidimensional as it uses more than one criterion to define gentrification. We show in Appendix O.2 that considering income and education jointly is important as both capture different spatial and temporal patterns of change. The downside of $M1$ is that it is discrete and relies on thresholds that are essentially arbitrary. This problem is compounded by the decadal nature of the census data we work with. Given identical starting points and income growth processes, blocks may exhibit gentrification or not depending on the starting date of the process.

We thus also construct two continuous and unidimensional measures: $M2$ is the per capita income growth in a poor block b over the decade, whereas $M3$ is the growth in the share of highly educated residents over a decade. These measures may capture less well the finer aspects of gentrification but they do not rely on arbitrary thresholds and are less sensitive to the dynamics of gentrification that may

[9]Since measuring variations is not meaningful when numbers are small, we also impose a minimum population threshold of 8 for the number of residents in the blocks.

not coincide with the decadal census data.

In estimating equation (1), we consider gentrification within the neighborhood around block b for each of the three definitions of a neighborhood g. For definition $M1$, we build a dummy variable equal to one if there is at least one poor block in the neighborhood that gentrifies according to $M1$, and zero otherwise. For definition $M2$ and $M3$, we aggregate by taking respectively the population-weighted average of per capita income log change and the population-weighted average change in the share of highly educated residents observed in poor blocks around a given block.[10]

3 Results

In this section, we first present results from estimating equation (1) and show that there is a robust set of pioneer sectors associated with gentrification. These sectors are identified using the gentrification episodes observed in New York over the decade 1990–2000. We then show that the presence of these pioneers correlates with gentrification, even when controlling for the usual determinants. These results are obtained by looking at gentrification episodes in New York and Philadelphia over the subsequent decade 2000–2010 and they are robust to using an instrumental variable (IV) strategy.

3.1 Who Are the Pioneers?

We estimate equation (1) separately for 429 6-digit NAICS industries in the New York urban area for $t = 1990$ and $\Delta t = 10$, and restrict the sample to blocks within

[10]Online appendix O.10 presents descriptive statistics on the geography of gentrification according to these three measures.

a 30 km radius around Wall Street.[11] We estimate 9 regressions per industry, one for each combination of the three measures of gentrification ($M1$, $M2$, and $M3$) and the three geographies ($g \in \{cont, 250m, 500m\}$). For each sector, we construct a score which reflects the number of specifications for which it exhibits atypical location patterns ($\widehat{\alpha}_1^s < 0$ and $\widehat{\alpha}_2^s > 0$, significant at the 1% level). A score of 9 means establishments in that sector are pioneers in each of the 9 regressions, whereas a score of 0 means establishments in that sector are never pioneers, no matter the measure of gentrification and the definition of neighborhoods.

Table 1 reports the distribution of scores. It shows that 79% of sectors cannot be considered as pioneers in any of these 9 specifications. At the other end of the spectrum, 26 sectors (6%) are pioneers in at least two-thirds of the specifications. Among them, eight sectors are pioneers in every specifications.

[Table 1 here]

In what follows, we focus on sectors with a minimum score of 6 as the number of sectors per score-value decreases significantly from 6 onward.[12] Hence, pioneer sectors are those identified as such in at least two-thirds of our specifications. The complete list of pioneer sectors is provided in Table 2.

[Table 2 here]

[11] There are 792 6-digit NAICS sectors with establishments in New York in 1990. For the estimates to be meaningful, we exclude sectors that are active in less than 100 blocks. We also keep only blocks with at least one establishment in any sector, i.e., we remove all purely residential blocks.

[12] We considered a more continuous measure of pioneers that uses the value of the scores. Yet, this turns out to be problematic when aggregating. It is, for example, not clear that 9 establishments with a score of 1 are equivalent to one establishment with a score of 9. If we aggregate up weighted counts of pioneers, we are likely to pick up locations with a large mass of establishments having each low scores.

Using the terminology introduced by Grodach et al. (2014), pioneer sectors are mostly found among commercial arts (Motion Picture and Video Production; Architectural Services; Industrial Design Services; Commercial Photography etc.) and fine arts (Independent Artists, Writers, and Performers; Art Dealers; Museums; Fine Arts Schools etc.) industries. Two pioneer sectors can be considered as consumer amenities (Full-Service Restaurants; All Other Specialty Food Stores). Solely—at the bottom of the list with a score of six—we find some sectors that are not related to arts and creativity in a broad sense, such as employment placement agencies.

Many of these pioneer sectors are mentioned frequently in newspaper articles and case studies on gentrification.[13] This list further shows creative industries are consistently found among the pioneers. This aspect is in line with the literature in the social sciences which emphasizes that gentrification is not only related to the growth of average income but also to the share of highly educated residents, the later not being necessarily wealthy.

[Table 3 here]

We checked the robustness of our list to alternative ways of estimating equation (1). Table 3 shows it is stable if we use the median per capita income log change and the median change in the share of highly educated residents in poor surrounding blocks instead of the population-weighted average. To account for the fact that there are so many zeroes in the count of establishments at the 6-digit NAICS and block level, we also estimate a logit model where the dependent variable is a dichotomous variable identifying, for a given 6-digit sector, the blocks with at least

[13] One example is art galleries or art dealers (NAICS 453920). See Schuetz (2014) and Easterly et al. (2018) for case studies. See also Grodach et al. (2014) on the association between commercial art industries and gentrification.

one establishment from that sector. As shown in the second column of Table 3, the list of pioneers obtained from this alternative specification is very similar to the one obtained from the negative binomial specification.

Instead of imposing two conditions to define pioneers ($\widehat{\alpha}_1^s < 0$ and $\widehat{\alpha}_2^s > 0$), we use only the second one. This leads to consider as pioneers all sectors that are overrepresented in poor neighborhoods that will subsequently gentrify. As shown in the third column, there are only six new sectors in the list, five of which are also related to the creative industries. Finally, when we run the analysis at the block-group level, at the level of which most census variables are available (see online Appendix O.8 for additional details), only seven sectors appear as pioneers but they are all part of the base list in Table 2. The fact that the list is much shorter when we use bigger geographic units such as block groups illustrates the fact that the mechanisms through which some pioneer sectors correlate with gentrification might be very local. Their identification then requires a finer geography than the one generally used in studies on gentrification (block groups or census tracts).

3.2 Pioneers Herald Gentrification

Pioneers are overrepresented in areas that subsequently gentrify. Yet, the gentrification process could be driven by other initial characteristics. In particular, demographic characteristics of the population, crime rates, the age of the housing stock, or various amenities may drive both the initial location of pioneer businesses and the subsequent arrival of more affluent and educated residents. We thus run a number of regressions to show that the presence of pioneer sectors remains significantly associated with gentrification even when controlling for these covariates. These results confirm that pioneers have an informational content that is not subsumed by these other variables.

To put distance between the data used to detect pioneers (New York between 1990 and 2000) and the data used to predict gentrification, we report results for a different decade (New York between 2000 and 2010) and for a different decade and city (Philadelphia between 2000 and 2010). We estimate the following equation:

$$\text{gentri}_{b,2000-2010} = \alpha + \gamma \cdot \Delta\text{pioneers}^{g}_{b,1990-2000} + (X^{g}_{b,2000})'\beta + \varepsilon^{g}_{b} \tag{2}$$

where $\text{gentri}_{b,2000-2010}$ is one of the three measures of gentrification for block b between 2000 and 2010; $\Delta\text{pioneers}^{g}_{1990-2000}$ is the change in the number of pioneer establishments over the same period; $X^{g}_{b,2000}$ is a set of controls measured in 2000 for neighborhood g of block b; and ε^{g}_{b} is an error term. We restrict our analysis to poor blocks.

Let $\mathcal{N}^{g}_{b}$ denote the set of neighboring blocks for measure $g \in \{cont, 250m, 500m\}$. We estimate equation (2) controlling for five types of initial characteristics (see the data appendix for details on the datasets we use and the construction of the variables). First, we control for socio-economic variables such as population and the neighborhood's racial composition. These variables are directly aggregated from b's neighboring blocks $j \in \mathcal{N}^{g}_{b}$. Second, we control for housing characteristics using rents in the block and the age of buildings in the neighborhood (computed over the blocks $j \in \mathcal{N}^{g}_{b}$, where we weight by the block's number of housing units). Third, we include several proxies for amenities: distance to parks; distance to public transportation (distance to the closest bus and train stop, number of lines in the area); a dummy for waterfront blocks; a count of major landmarks and—for the subsample of blocks in the five boroughs of New York City—indicators for limited height and historical districts and for the number of property crimes and of violent crimes.[14]

[14] Although the NYPD and the City of New York have made finely grained geographic data on crime available, the public release only goes back to 2018. We rather use the historic publicly available data

Fourth, for the sample restricted to New York City again, we include controls related to the presence of rent controlled buildings. Last, we control for the spatial diffusion of gentrification by including a proxy for the presence of gentrifying blocks in the neighborhood during the preceding decade, 1990–2000.[15] Count variables like the number of bus lines, metro lines, and landmarks are measured as $\ln(1+\sum_{j\in\mathcal{N}_b^g}\text{number}_j)$. Because some variables are aggregated or averaged over the neighborhoods, we may have spatial correlation in the data in (2). To correct for this, we use the HAC method proposed by Conley (1999) and the Stata package made available by Fabrizio et al. (2018).

The coefficient of interest is γ, which measures the relationship between a change in the number of pioneer businesses around block b between 1990 and 2000, and subsequent socio-economic changes (gentrification) in that block between 2000 and 2010. A positive estimate of γ means that, conditional on our controls, block b experienced stronger socio-economic changes if it was located in an area that saw a larger increase in the number of pioneer establishments. The change in the number of pioneers is constructed using the base list of pioneer sectors detected in subsection 3.1 and summarized in Table 2. Let $\mathcal{P}$ denote the set of pioneer sectors. The measure is defined as

$$\Delta\text{pioneers}^g_{b,1990-2000} \equiv \sum_{s\in\mathcal{P}}\sum_{j\in\mathcal{N}_b^g} n^{s,j}_{b,2000} - \sum_{s\in\mathcal{P}}\sum_{j\in\mathcal{N}_b^g} n^{s,j}_{b,1990} \tag{3}$$

that are available at the precinct level, precincts being bigger units than blocks. We thus assign the precinct information to the blocks belonging to that precinct.

[15] When the dependent variable is the discrete measure of gentrification $M1$, we include the log of the distance to the closest block that gentrified in the previous decade. For the continuous measures of gentrification $M2$ and $M3$, the corresponding controls are the neighborhood change in log per capita income or in the share of educated residents between 1990 and 2000.

where $n_{b,t}^{s,j}$ is the number of establishments from pioneer sector s in block j belonging to neighborhood g around block b in year t. Observe that our measure can be positive or negative depending on net entry or exit of pioneers in $\mathcal{N}_b^g$.

Panel (a) of Table 4 summarizes the results obtained for New York and New York City and for neighborhoods defined with a 500 meter radius around blocks.[16] It shows that for all three measures of gentrification, there is a positive and statistically significant relationship between the growth in the number of pioneers between 1990–2000 and gentrification between 2000–2010, conditional on the controls. Hence, pioneers contain information not subsumed by the other covariates. It is worth noting that we also control for the change in the number of non-pioneer establishments in the neighborhood around the block in our regressions. Strikingly, Table O.8 in the online appendix shows that it is only changes in pioneer businesses that are positively correlated with subsequent gentrification episodes. Indeed, the coefficient on the change in the number of other (non-pioneer) establishments is almost never significantly positive.

[16]We report estimates of our controls in the online appendix (see Table O.8). In line with previous studies, there are no clear patterns regarding the impact of racial composition on subsequent gentrification (McKinnish et al., 2010). Turning to housing characteristics, buildings around gentrifying blocks tend to be older. This is in accord with the literature that emphasizes the importance of neighborhood housing cycles and filtering induced by the deterioration and subsequent renewal of the housing stock (Rosenthal, 2008; Brueckner and Rosenthal, 2009). Transport access matters little for NYC, but we find precisely estimated positive effects of the number of train lines for the larger metro area, which suggests access to transportation is valued more in the outlying parts of the city. Consistent with Guerrieri et al. (2013), we also find evidence of spatial diffusion. All else being equal, the closer a block is to a block that gentrified in the previous decade, the higher the probability that it also gentrifies subsequently, at leat in the largest metro area.

3.3 IV Regressions

Our estimates do not necessarily capture the causal effect of pioneers on gentrification since unobserved local shocks may drive both the change in the presence of pioneers between 1990 and 2000, and the demographic change of a block between 2000 and 2010.[17] Because we have no quasi-experimental variation to exploit, we construct an IV for the change in the number of pioneer establishments. We first predict the number of pioneers in each block in 2000 by combining the initial stock of pioneers in the neighborhood in 1990 with the U.S. national growth rates (excluding New York) in the number of establishments in pioneer industries, and we then subtract the initial number of pioneers in the neighborhood. Formally, let

$$\widehat{n}^{s,g}_{b,2000} \equiv \sum_{j \in \mathcal{N}^g_b} n^{s,j}_{b,1990} \times \left(\Delta n^{s,\text{US}}_{1990-2000} \right) \%, \tag{4}$$

so that

$$\Delta \text{pioneersIV}^g_b \equiv \sum_{s \in \mathcal{P}} \sum_{j \in \mathcal{N}^g_b} \left(\widehat{n}^{s,j}_{b,2000} - n^{s,j}_{b,1990} \right) \tag{5}$$

is our instrument.[18] Its validity relies on two identification assumptions, the relevance of which cannot be easily assessed from a statistical viewpoint.

First, the local dynamics of gentrification is unrelated to the national growth of pioneer industries. We believe this assumption is the least questionable: though New York is the largest metropolitan area in the U.S., it is unclear how a local shock there may drive the growth of pioneers industries in the rest of the country. Note that to

[17] If these local shocks also induce demographic changes between 1990 and 2000, they are captured by the lagged demographic changes included in the specifications.

[18] We present the instrument computed from information on the presence of pioneers in a 500m radius. The instrument does not perform well for smaller radii, because the evolution of the presence of pioneers at a very local level is extremely hard to predict.

make sure that local changes in New York do not affect the global dynamics of these industries, we compute the growth of these industries in the U.S. excluding New York.

Second, local shocks driving the demographic changes of a block between 2000 and 2010 do not drive the level of pioneers in 1990. This assumption is violated if pioneers' presence in 1990 is correlated with block-level gentrification shocks between 2000 and 2010. For instance, it could be the case that in 1990 pioneers choose a block based on unobserved amenities (e.g., the presence of warehouses that can be turned into work spaces or industrial lofts), and these amenities then become fashionable in the 2000s for wealthy or highly educated people. We control for a range of amenities that may influence the presence of pioneers (housing prices, limited height districts, historical districts, distance to parks, etc.), but some of them might be unobserved like the presence of industrial buildings. What is reassuring is that, although emblematic, industrial buildings are not present in all gentrified neighborhoods (think of Harlem, for instance).[19]

[Table 4 here]

Panel (b) of Table 4 shows that the IV coefficients are smaller than the OLS coefficients. Yet, reassuringly, the results are quite similar to the OLS estimates once standard errors are accounted for. In a nutshell, there seems to be no substantial endogeneity bias in our baseline estimations. Note that the results remain interesting even if our identification assumptions are violated. In that case, the mere presence

[19] Another channel are housing cycles (Brueckner and Rosenthal, 2009; Rosenthal and Ross, 2015). Pioneers could be attracted in 1990–2000 by neighborhoods with old and cheap real estate, which becomes obsolete in 2000–2010 and gets replaced with new (or substantially renovated) housing. The latter is known to attract more affluent residents. We thus control for the initial age of the housing stock.

of pioneers is not a cause *per se* of a neighborhood's gentrification. Yet pioneers still herald gentrification in the sense that their presence systematically signals the unobserved local shocks driving gentrification that are not captured by the "usual suspects" of gentrification.

3.4 Robustness Checks

As a first robustness check, panels (c) and (d) of Table 4 report OLS estimates for New York using a tighter definition of neighborhoods (250 meters radius or contiguous blocks). As one can see, the results are robust. If anything, the effects of pioneers become stronger, thereby suggesting that the association between pioneer establishments and gentrification is tighter the more local the analysis is.

As a second robustness check, we extend our analysis in Table 5 by looking at gentrification in the Philadelphia metropolitan area between 2000 and 2010 (see Table O.9 in the online appendix for a full set of estimates). We estimate equation (2) by both OLS and IV using, as before, the list of pioneers identified from the New York data between 1990 and 2000. As for New York, we build a stable geography for Philadelphia and capture gentrification over the period 2000–2010. Using a different city and a different decade than those used to identify pioneers arguably puts additional distance between the identification of pioneers and the assessment of their predictive power.

The results in Table 5 are very similar to those obtained for New York: blocks more exposed to changes in the number of pioneers between 1990 and 2000 are more likely to gentrify over the subsequent decade. If anything, the results seem even stronger for Philadelphia.[20] Since we use in our estimations a different city and a

[20]There is more variation in block-level socio-demographic changes in Philadelphia between 2000 and 2010 than in New York, which may explain the stronger effects. We also estimated the effects for

different period than those used to identify the pioneers, we view these results as a strong vindication of the association between pioneers and subsequent neighborhood change, irrespective of whether that association reflects causality or correlation.

[Table 5 here]

4 Why Pioneers Matter for Gentrification

We have identified a set of sectors—mostly artistic and creative—that are statistically overrepresented in soon-to-gentrify neighborhoods. The goal of this section is to discuss why their presence provides information regarding future gentrification. In this discussion, we disentangle causal mechanisms—in which pioneer establishments or workers play an active role—from alternative explanations—a simple correlation between the presence of pioneers and unobserved drivers of gentrification.

4.1 Effects of Pioneer Establishments

To play an active role in the gentrification process, pioneers should be catalysts for the arrival of wealthy and educated residents in initially poor neighborhoods. The location of pioneers establishments in a poor neighborhood may change the demography of a neighborhood through two main channels. First, pioneers may provide consumption amenities valued by wealthy and educated residents. Some of the pioneer industries we have identified provide such amenities (Full-Service restaurants; All Other Amusement and Recreation Industries; All Other Specialty Food Stores; Museums). Their presence may compensate for other disamenities of poorer neigh-

Boston but were unable to find strong results. One key reason is that the distributions of block-level income and education changes are much more uniform in Boston which already largely gentrified before 2000.

borhoods and attract residents that usually locate in wealthier places. This view is consistent with Baum-Snow and Hartley (2016) and Couture and Handbury (2020) who find that local amenities increasingly valued by young educated people explain part of the urban revival observed in U.S. central cities.

Beyond this direct impact, amenities provided by pioneers may also interact with other amenities appreciated by wealthy and educated residents. In particular, we test whether the relationship between gentrification and the change in the number of pioneers varies with the distance to large high-skilled employment centers (prime locations; Ahlfeldt et al. 2020). Stronger positive effects close to prime locations would suggest that pioneer businesses may be interacting with skilled workers' desire to reduce commute time, a channel put forward in the literature (e.g., Edlund et al., 2015; Brown et al., 2016). As shown in Table O.10 in the online appendix, there are only weak and insignificant interaction effects between local pioneer businesses and distance to prime locations, which is consistent with Couture and Handbury (2020) who find that distaste for commuting does not explain much of the recent demographic dynamics of urban centers. We also look at interactions with natural amenities as measured by distance to the waterfront, or to open water more generally. The interaction between pioneers and distance to waterfront is negative and significant in most specifications, although the estimated coefficient is small. The link between the presence of pioneers and gentrification is thus only slightly reinforced by the proximity to open water.

[Figure 1 here]

Pioneers may also be a catalyst for gentrification because their presence is a reassuring signal for wealthy residents who hesitate to move to a cheaper neighborhood. This mechanism prevails if prospective residents believe that architects, designers, or

artists have a better anticipation of the future prospects of a neighborhood. For those signals to matter, pioneers should be visible to prospective newcomers, e.g., when walking through a neighborhood. To gauge the visibility of pioneers, we tentatively computed the distribution of establishments across floors, separating pioneers from non-pioneer establishments. Figure 1 reveals that a fourth of pioneer establishments are located within the first three floors of their buildings and are, therefore, likely to be visible from the street. Put differently, pioneer establishments are relatively visible in a neighborhood, which may support the idea that they provide signals to future residents as to the potential upside of the neighborhood.

4.2 Effects of Pioneer Establishments' Workers

Whereas the previous section emphasizes how the location of pioneer establishments *per se* may influence future gentrification, we now argue that specific traits of the workforce of these pioneers may also explain their role in the gentrification process. We first examine the socio-demographic profile of workers employed by these sectors. We use IPUMS microdata for the years 2000 and 2010 and compute a set of worker characteristics in the New York metropolitan area, distinguishing pioneer from non-pioneer sectors.[21]

Figure 2 shows systematic differences in the characteristics of workers employed by pioneer and non-pioneer sectors. First, as panels (a) and (b) show, these sectors employ somewhat younger but substantially more educated workers. However, wages in pioneer sectors do not markedly differ from wages in the other sectors, which suggests again that income and education play subtly different roles in the gentrification process and should both be considered. Panels (d) to (f) summarize

[21] Results using data for the entire U.S. are qualitatively similar, though less marked. They are relegated to Figure O.4 in the online appendix.

the most striking differences, which are all related to household composition: workers in pioneer industries are more often single, more often in power couples (i.e., both members are college-educated), and they have fewer children. Finally, panels (g) and (h) show that workers employed in pioneer sectors tend to work closer to their place of residence: they work more often at home, and they commute more often by bicycle or by foot than workers in non-pioneer sectors.

The big picture that emerges is that workers in pioneer industries tend to live closer to their workplace, which may create a link between the presence of pioneer establishments and the socio-demographic composition of the neighborhood. In addition, all the aforementioned socio-demographic characteristics of pioneers' workers (age, education, marital status, power couples) are linked to the type of population that is usually associated with gentrification of urban neighborhoods. For instance, recent evidence suggests that: (i) urban revival has been partly driven by young educated millennials (Baum-Snow and Hartley, 2016; Couture and Handbury, 2020); (ii) these millennials work with more flexible employment relationships than previous generations (Aguiar et al., 2017); and (iii) these millennials have a different travel behavior and a distaste for commuting (Edlund et al., 2015; Brown et al., 2016). When workplace and residence become more tightly connected—since workers want to spend less time commuting—the presence of pioneer businesses goes hand-in-hand with the local presence of "pioneer residents", which might explain the relationship between the mix of businesses in a given area and its subsequent evolution in terms of residents.

The location of creative workers close to their employers may also be an extra factor of attractiveness of poor neighborhoods hosting pioneer establishments. Indeed, one conjecture in urban studies is that non-creative high-income households value the geographic proximity to artists (Ley, 2003; Grodach et al., 2014). Pioneers thus

increase the stock of 'cultural capital' in a neighborhood, which acts as an amenity that may attract future wealthy and educated residents.

[Figure 2 here]

Last, the initial move of pioneer establishments and their workers to some neighborhood may simply reflect changes in preferences. For instance, the growing interest of upper classes in street culture may have occurred first in artistic circles (in which workers of pioneer establishments operate), and then become mainstream among other wealthy and educated people. Similarly, the initial move of pioneer workers to some neighborhoods may reflect a change in tastes for certain types of architecture. The location of pioneer establishments may thus reflect early changes in preferences for some traits of poor neighborhoods. In such a case, pioneer establishments do not have a causal impact on gentrification but their presence reveals 'deep drivers' of gentrification that are not readily observed by researchers.

4.3 The singularity of pioneers

The pioneers identified in Section 3.1, and their differences with other sectors uncovered in Section 4.2, lead to two main observations. First, many pioneers are in creative industries. Second, pioneers are businesses that tend to employ young and educated workers who live 'next to their job'. Two legitimate questions are: (i) why not all creative industries are pioneers; and (ii) whether some non-pioneer industries that employ workers with similar characteristics as pioneer industries would have the same impact.

To answer these questions, we build two additional lists of sectors. The first gathers all sectors that are creative but not identified as pioneers. To select these sectors, we first consider all 6-digit NAICS categories that belong to the same 4-digit NAICS

categories as our pioneers. Most of these non-pioneer sectors are not creative.[22] We nonetheless find eleven 6-digit non-pioneer sectors that can be considered as creative. They include "theater companies", "photography studios", "landscape architectural services", or "music publishers". We also ran a lexicographic search with the words "art", "artists", "design". Last, we manually checked for other sectors that we judge as creative. We end up with 19 sectors, which are displayed in column (a) of Table 6.

The second list gathers sectors which are neither pioneer nor creative, but whose workers have characteristics close to those of workers in pioneer sectors. More specifically, we tag sectors whose workers' mean age is below the third quartile of the mean age of pioneers' workers; whose share of educated is above the first quartile of pioneers' share of educated; and whose share of workers working close to their home is above the first quartile of that for pioneers' workers. We end up with 12 sectors presented in column (b) of Table 6 including "veterinary services", and a variety of consulting, scientific, and technical services.

[Table 6 here]

Table 7 compares pioneer sectors with sectors in these two lists. Column (2) shows that pioneer sectors are indistinguishable from non-pioneer creative sectors in terms of worker characteristics. They differ, however, along two important dimensions. First, the average number of employees per establishment is smaller for pioneer than for non-pioneer creative industries (8 *vs* 26 employees). Second, there are many more establishments in pioneer sectors than in non-pioneer creative sectors (76K *vs* 19K establishments per sector on average). The difference in the number of establishments

[22] For instance, "museums" is in the same 4-digit category as "nature park"; "fine art schools" is in the same 4-digit category as "automobile driving schools"; "commercial photography" is in the same category as "veterinary services"; and "art dealer" is in the same category as "tobacco stores".

between the two might signal that non-pioneer sectors are more 'capital intensive' or need more infrastructure (R&D labs, colleges and universities, libraries, or theater companies require more infrastructure than photo studios or art galleries). The fact that establishments in pioneer sectors are smaller and more footloose may explain why they are found more systematically in gentrifying neighborhoods: they can relocate more quickly and exploit the opportunities offered to them by some poorer neighborhoods.

For the non-pioneer sectors employing workers with similar traits (Column (3) of Table 7), wages are the main dimension along which they differ markedly. Workers with similar characteristics that do not work in pioneer or other creative sectors earn more and are more likely to be in a power couple. The wage differential may explain why these other sectors are not present in poor neighborhoods: their workers may want to work close to their home, but can afford to live in more affluent areas.

[Table 7 here]

In Table O.11 in the online appendix, we have introduced the establishments belonging to these extra-lists as additional determinants of gentrification. Only pioneers are positively associated with subsequent gentrification. If anything, the non-creative non-pioneer industries with similar workers and the non-pioneer creative sectors are negatively associated with subsequent gentrification.

Overall, our results suggest that pioneer establishments and their workers are more willing and able to go to poorer neighborhoods. Pioneer workers' decisions might be driven by pecuniary reasons, because they earn lower wages than workers with similar characteristics working in non-creative sectors. It might also be the case that pioneer workers embrace a lifestyle that trades off differently the current

disamenities and the future prospects these deprived neighborhoods might offer.[23] The fact that pioneers are mostly small businesses that are more footlose may also explain why they can take the risk to experiment with more uncertain areas that offer a higher expected potential.

5 Conclusion

Our analysis improves our understanding of where gentrification occurs within a city by focusing on the specific role of businesses. We make two contributions to the literature on gentrification. First, we propose a method to detect sectors that are usually found in more affluent neighborhoods but are overrepresented in poorer areas that will experience future gentrification. These 26 pioneer sectors subsumes mostly cultural and creative industries, especially 'commercial arts'. To the best of our knowledge, it is the first set of sectors associated with gentrification derived from an econometric analysis using micro-data.

Second, we show that the presence of pioneer businesses has explanatory power for future gentrification beyond a large set of controls already pointed out in the literature. This effect survives a battery of tests and holds when applied to another city not used to identify pioneers.

Last, we discuss several mechanisms through which the presence of pioneers and their workers may influence gentrification. Some of these mechanisms imply a causal role of pioneers in the gentrification process, consistent with the IV estimates we present. Without relying on the IV and on a causal interpretation of our results, the presence of pioneers may still capture drivers of gentrification that would be otherwise unobserved. Providing firmer evidence for (or against) these mechanisms is

[23] See Ley (2003) for a perspective about creative workers, neighborhood aesthetics, cultural capital, and gentrification.

a promising avenue for future research. We also think that these sectors can help to build indexes of the gentrification prospects of neighborhoods, which should be useful to researchers, practitioners, and policy makers concerned with these dynamics.

Acknowledgements. We thank the three referees who reviewed this paper and the editor Amit Khandelwal for their very helpful comments. We further thank our discussants Felipe Carozzi, David Cuberes, Amy Schwartz, Coen Teulings, and Victor Ye, as well as Bocar Ba, Nate Baum-Snow, Victor Couture, Jessie Handbury, Rachel Meltzer, Yasusada Murata, Amine Ouazad, Seyhun Sakalli and participants at various conferences and seminars for constructive comments and suggestions. We especially thank Stuart Rosenthal for sharing his floor-level data with us. Alek Racicot provided able research assistance. We gratefully acknowledge financial support from the SSHRC Insight Grants program ('Cities in motion', grant #435-2016-1246). Behrens gratefully acknowledges financial support from the CRC Program of SSHRC for the funding of the Canada Research Chair in Regional Impacts of Globalization. We acknowledge the support of the HSE University Basic Research Program. Martin acknowledges financial support from UQAM for the funding of the UQAM research chair on the local impact of multinational firms. Any remaining errors are ours.

The views expressed in this paper are those of the authors and do not necessarily represent the views of Statistics Canada or the Government of Canada.

References

Aguiar, Mark, Mark Bils, Kerwin Kofi Charles, and Erik Hurst, "'Leisure luxuries and the labor supply of young men'," NBER Working Papers 23552, National Bureau of Economic Research, Inc June 2017.

Ahlfeldt, Gabriel M., Thilo N.H. Albers, and Kristian Behrens, "Prime locations," CEPR Discussion Paper DP15470, Centre for Economic Policy Research November 2020.

Barton, Michael, "An exploration of the importance of the strategy used to identify gentrification," *Urban Studies*, 2016, *53* (1), 92–111.

Baum-Snow, Nathaniel and Daniel Hartley, "'Accounting for central neighborhood change, 1980-2010'," Working Paper Series WP-2016-9, Federal Reserve Bank of Chicago September 2016.

Brown, Anne, Evelyn Blumenberg, Brian Taylor, Kelcie Ralph, and Carole Voulgaris, "A taste for transit? Analyzing public transit use trends among youth," *Journal of Public Transportation*, 03 2016, *19* (1), 49–67.

Brueckner, Jan K. and Stuart S. Rosenthal, "Gentrification and neighborhood housing cycles: Will America's future downtowns be rich?," *The Review of Economics and Statistics*, November 2009, *91* (4), 725–743.

Brummet, Quentin and Davin Reed, "'Gentrification and the well-being of original neighborhood residents: Evidence from longitudinal census microdata'," Mimeo 2018.

Conley, T. G., "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Costa, Dora and Matthew Kahn, "Power couples: Changes in the locational choice of the college educated, 1940-1990," *Quarterly Journal of Economics*, 2000, *115* (4), 1287–1315.

Couture, Victor and Jessie Handbury, "'Urban revival in America'," *Journal of Urban Economics*, 2020, *119* (103267).

_ , **Cecile Gaubert, Jessie Handbury, and Erik Hurst**, "'Income growth and the distributional effects of urban spatial sorting'," Mimeo 2018.

Ding, Lei, Jackelyn Hwang, and Eileen Divringi, "Gentrification and residential mobility in Philadelphia," *Regional Science and Urban Economics*, 2016, *61* (C), 38–51.

Easterly, William, Laura Freschi, and Steven Pennings, "'A long history of a short block: Four centuries of development surprises on a single stretch of a New York City street'," Mimeo, NYU Development Research Institute and World Bank 2018.

Edlund, Lena, Cecilia Machado, and Maria Micaela Sviatschi, "'Bright minds, big rent: Gentrification and the rising returns to skill'," Working Paper 21729, National Bureau of Economic Research November 2015.

Ellen, Ingrid Gould, Keren Mertens Horn, and Davin Reed, "Has falling crime invited gentrification?," *Journal of Housing Economics*, 2019, *46* (101636), 1101636.

Fabrizio, Colella, Rafael Lalive, Seyhun Orcan Sakalli, and Mathias Thoenig, "'Inference with arbitrary clustering'," Mimeo, HEC University of Lausanne 2018.

Freeman, Lance and Frank Braconi, "Gentrification and displacement New York City in the 1990s," *Journal of the American Planning Association*, 2004, *70* (1), 39–52.

Glaeser, Edward L. and Matthew E. Kahn, "Decentralized employment and the transformation of the American city," *Brookings-Wharton Papers on Urban Affairs*, 2001, pp. 1–63.

_ , **Hyunjin Kim, and Michael Luca**, "Nowcasting gentrification: Using Yelp data to quantify neighborhood change," *AEA Papers and Proceedings*, 2018, *108* (5), 77–82.

Grodach, Carl, Nicole Foster, and James Murdoch III, "Gentrification and the Artistic Dividend: The Role of the Arts in Neighborhood Change," *Journal of the American Planning Association*, 2014, *80* (1), 21–35.

Guerrieri, Veronica, Daniel Hartley, and Erik Hurst, "Endogenous gentrification and housing price dynamics," *Journal of Public Economics*, 2013, *100* (C), 45–60.

Kan, Kamhon, Sunny Kai-Sun Kwong, and Charles Ka-Yui Leung, "The dynamics and volatility of commercial and residential property prices: Theory and evidence," *The Quarterly Journal of Economics*, 2004, *44* (1), 95–123.

Lees, Loretta, "Super-gentrification: The case of Brooklyn Heights, New York City," *Urban Studies*, 2003, *40* (12), 2487–2509.

Lester, T. William and Daniel A. Hartley, "The long term employment impacts of gentrification in the 1990s," *Regional Science and Urban Economics*, 2014, *45* (C), 80–89.

Ley, David, "Artists, Aestheticisation and the Field of Gentrification," *Urban Studies*, 2003, *40* (12), 2527–2544.

Liu, Crocker H., Stuart S. Rosenthal, and William C. Strange, "The vertical city: Rent gradients, spatial structure, and agglomeration economies," *Journal of Urban Economics*, 2018, *106*, 101 – 122.

McKinnish, Terra, Randall Walsh, and T. Kirk White, "Who gentrifies low-income neighborhoods?," *Journal of Urban Economics*, 2010, *67* (2), 180–193.

Meltzer, Rachel, "Gentrification and small business: Threat or opportunity?," *Cityscape*, 2016, *18* (3), 57.

_ **and Pooya Ghorbani**, "Does gentrification increase employment opportunities in low-income neighborhoods?," *Regional Science and Urban Economics*, 2017, *66*, 52 – 73.

O'Sullivan, Arthur, "Gentrification and crime," *Journal of Urban Economics*, January 2005, *57* (1), 73–85.

Rosenthal, Stuart S., "Old homes, externalities, and poor neighborhoods. A model of urban decline and renewal," *Journal of Urban Economics*, May 2008, *63* (3), 816–840.

_ **and Stephen L. Ross**, "Chapter 16 - Change and persistence in the economic status of neighborhoods and cities," in J. Vernon Henderson Gilles Duranton and William C. Strange, eds., *Handbook of Regional and Urban Economics*, Vol. 5 of *Handbook of Regional and Urban Economics*, Elsevier, 2015, pp. 1047 – 1120.

Schuetz, Jenny, "Do art galleries stimulate redevelopment?," *Journal of Urban Economics*, 2014, *83*, 59 – 72.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick, "Capitalists in the Twenty-first Century," *The Quarterly Journal of Economics*, 2019, *134* (4), 1675–1745.

Su, Yichen, "'The rising value of time and the origin of urban gentrification'," Mimeo 2018.

Sullivan, Daniel Monroe and Samuel C. Shaw, "Retail gentrification and race: The case of Alberta street in Portland, Oregon," *Urban Affairs Review*, 2011.

Zukin, Sharon, Valerie Trujillo, Peter Frase, Danielle Jackson, Tim Recuber, and Abraham Walker, "New retail capital and neighborhood change: Boutiques and gentrification in New York City," *City and Community*, 2009, *8* (1), 47–64.

Table 1: Distribution of scores across sectors.

Score	Number of sectors	Cumulative number
9	8	8
8	3	11
7	6	17
6	9	26
5	12	38
4	9	47
3	14	61
2	21	82
1	25	107
0	322	429

Notes: The score is the number of specifications in which a sector meets the criteria to be considered as a pioneer ($\widehat{\alpha}_1^s < 0$ and $\widehat{\alpha}_5^s > 0$ and significant at the 1% level in equation (1)). The maximum score is 9 as we estimate 9 distinct regressions using the pairwise combinations of the three measures of gentrification and the three definitions of a neighborhood (250m radius, 500m radius, or block contiguity).

Table 2: Pioneer sectors.

Score	Code	NAICS 6-digit sector	Score	Code	NAICS 6-digit sector
9	512110	Motion Picture and Video Production	7	541430	Graphic Design Services
9	541310	Architectural Services	7	541922	Commercial Photography
9	711130	Musical Groups and Artists	7	611610	Fine Arts Schools
9	711320	Promoters of Performing Arts and Similar Events without Facilities	7	712110	Museums
9	711410	Agents and Managers for Artists, and Other Public Figures	6	445299	All Other Specialty Food Stores
9	711510	Indep. Artists, Writers, and Performers	6	511130	Book Publishers
9	722110	Full-Service Restaurants	6	512240	Sound Recording Studios
9	813990	Other Membership Organizations	6	541330	Engineering Services
8	511120	Periodical Publishers	6	541410	Interior Design Services
8	541810	Advertising Agencies	6	541512	Computer Systems Design Services
8	713990	All Other Amusement and Recreation Industries	6	541618	Other Management Consulting Services
7	453920	Art Dealers	6	541820	Public Relations Agencies
7	541420	Industrial Design Services	6	561310	Employment Placement Agencies

Notes: List of all sectors identified as pioneers ($\widehat{\alpha}_1^s < 0$ and $\widehat{\alpha}_2^s > 0$ and significant at the 1% level in equation (1) in at least six out of the nine specifications). "Code" refers to the 6-digit code in the 2002 NAICS classification. Other Membership Organizations (NAICS 813990) comprises establishments primarily engaged in promoting the interests of their members (except religious organizations, social advocacy organizations, civic and social organizations, business associations, professional organizations, labor unions, and political organizations). It includes, for example, art councils, condominium owners' associations, regulatory athletic associations, property owners' associations, and sport leagues.

Table 3: Stability of the list of pioneer sectors.

Median instead of weighted average	Logit regressions	Alternative definition	Block-group level geography
In base list : 19 New in list: 0	In base list: 22 New in list: 4	In base list: 26 New in list: 6	In base list: 7 New in list: 0

Notes: We report 4 alternative sets of results, obtained from 4 alternative specifications: (i) using the median changes rather than the weighted average; (ii) using a dummy for the presence of a sector rather than the number of establishments (and thus a logit model rather than a negative binomial model); (iii) defining pioneers as sectors overrepresented in poor soon-to-gentrify neighborhoods irrespective of their probability to locate in poor neighborhoods; and (iv) performing the analysis at the block-group level rather than at the block level. 'In base list' reports the number of sectors that appear both in the alternative specification and in the baseline list. 'New in list' reports the number of sectors in the alternative specification that are identified as pioneers but do not appear in the baseline list. Recall that there are 26 pioneer sectors in the base list.

Table 4: Pioneers and gentrification in New York, 2000–2010.

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
	(a) Blocks within a 500-meter radius, OLS					
Δ # pioneer estab. (1990–2000)	1.389[a]	1.009[a]	1.778[a]	1.498[a]	0.399[a]	0.266[a]
	(0.311)	(0.244)	(0.323)	(0.273)	(0.124)	(0.090)
# of observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.047	0.072	0.055	0.078	0.035	0.072
	(b) Blocks within a 500-meter radius, IV					
Δ # pioneer estab. (1990–2000)	0.843[a]	0.483[c]	1.131[a]	1.020[a]	0.196[b]	0.144[c]
	(0.256)	(0.284)	(0.307)	(0.305)	(0.081)	(0.083)
# of observations	34,164	20,005	33,856	19,822	33,863	19,828
Kleinbergen-Paap F-stat	11.498	12.087	12.255	12.581	12.404	12.857
	(c) Blocks within a 250-meter radius, OLS					
Δ # pioneer estab. (1990–2000)	3.888[a]	1.857[b]	6.500[a]	4.837[a]	1.411[a]	0.480
	(0.878)	(0.884)	(0.914)	(0.942)	(0.380)	(0.391)
# of observations	34,164	19,867	33,844	19,679	33,853	19,687
R-squared	0.037	0.069	0.072	0.096	0.039	0.076
	(d) Contiguous blocks, OLS					
Δ # pioneer estab.	3.430[a]	1.560	5.563[a]	3.903[a]	1.141[b]	0.160
	(1.159)	(1.230)	(1.209)	(1.324)	(0.484)	(0.489)
# of observations	34,164	19,882	33,851	19,695	33,860	19,703
R-squared	0.033	0.066	0.070	0.095	0.033	0.073
Controls	✓	✓	✓	✓	✓	✓
Sample	New York	NYC	New York	NYC	New York	NYC

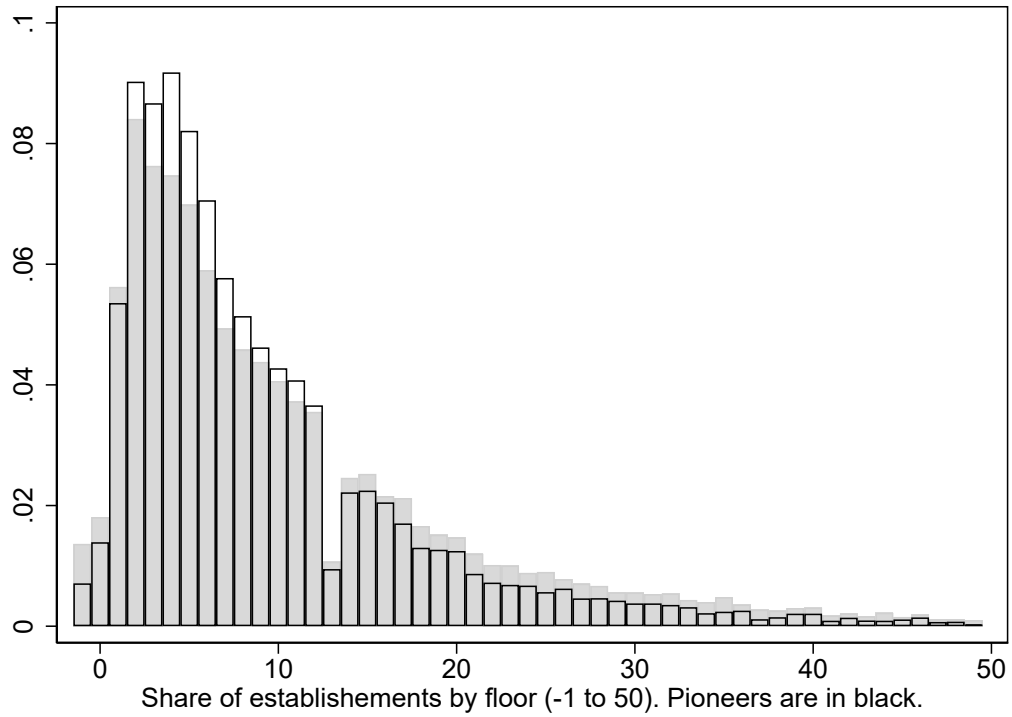
Notes: Reported coefficients and standard errors are mutilplied by 1,000 compared to the actual ones. All regressions include as controls: Ln per cap. income; Share college edu. resid.; Ln rent; Median age of buildings; Share black resid.; Share asian resid.; Share other resid.; Ln population; Less than 200m from waterfront; Ln (1+# train lines); Ln (1+# bus lines); Ln distance to closest park; Ln (1+# of main landmarks); Socio-economic changes in the neighborhood 1990–2000. For the sample limited to NYC, the controls also include # murder per cap.; # burglary per cap.; # robbery per cap.; # rape per cap.; Ln (1+# rent control buildings); Share vacant land; Presence of limited height districts; Presence of historical districts. See Table O.8 in the online appendix for the full results. All explanatory variables in levels are measured in 2000 and are computed using different definitions of neighborhoods (250 or 500 meters rings, or contiguous blocks), except for the distance variables and the waterfront dummy. Robust standard errors, corrected for cross-sectional spatial dependence (using HAC estimation), are reported in parentheses. [a] = significant at 1%, [b] = significant at 5%, [c] = significant at 10%.

Table 5: Pioneers and gentrification in Philadelphia, 2000–2010.

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
Δ # pioneer estab.	18.183^{a}	31.757^{a}	6.988^{a}	15.756^{c}	1.574^{a}	4.139^{b}
	(2.938)	(12.073)	(2.313)	(8.472)	(0.443)	(1.676)
Controls	✓	✓	✓	✓	✓	✓
# of observations	18,144	18,144	18,009	18,009	18,141	18,141
R-squared	0.157	n.a.	0.064	n.a.	0.089	n.a.
Specification	OLS	IV	OLS	IV	OLS	IV
Kleinbergen-Paap F-stat	n.a.	7.925	n.a.	7.929	n.a.	8.028

Notes: Reported coefficients and standard errors are multiplied by 1,000 compared to actual ones. All regressions include: Ln per cap. income; Share college edu. resid.; Ln rent; Median age of buildings; Share black resid.; Share asian resid.; Share other resid.; Ln population; Less than 200m from waterfront; Ln distance to subway; Ln distance to closest park; # of main landmarks; Socio-economic changes in the neighborhood 1990–2000. See Table O.9 in the online appendix for the full results. The measure of the change in the number to pioneers is given by equation (3) in the text. All explanatory variables in levels are measured in 2000 and are computed using 500m rings, except for the distance variables and the waterfront dummy. Robust standard errors, corrected for cross-sectional spatial dependence within a 500 meters radius (using HAC estimation), are reported in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

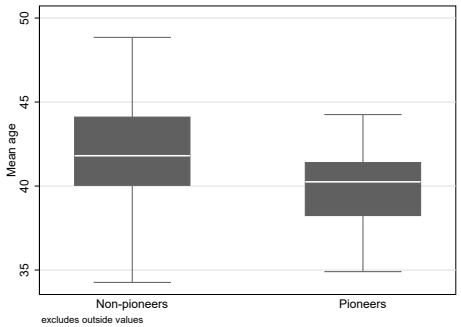
Figure 1: Distribution of establishments across floors, pioneers vs non-pioneers.



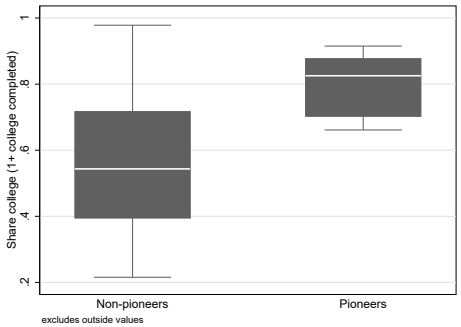
Notes: Computation based on D&B establishments located in Manhattan in 2018. See the data Appendix and Liu et al. (2018) for more details.

Figure 2: Selected characteristics of workers in pioneer sectors in New York, 2000–2010.

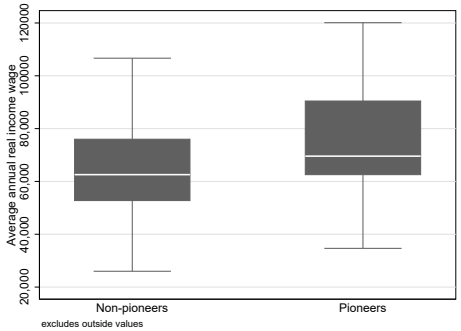
(a) are a bit younger.



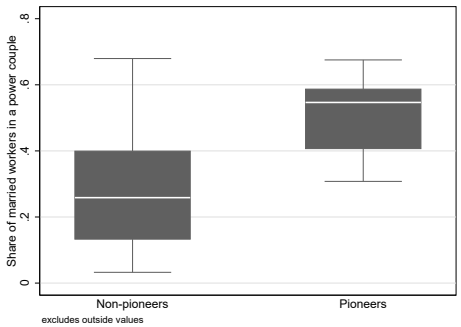
(b) are more educated.



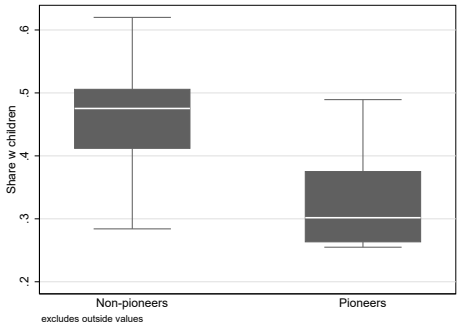
(c) are paid a bit more.



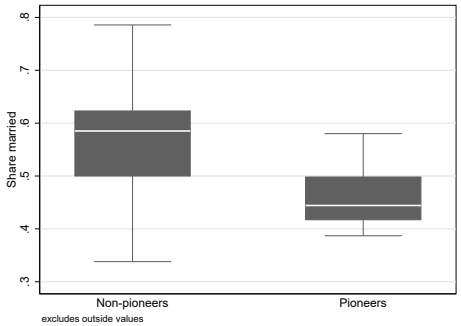
(d) are in 'power couples'.



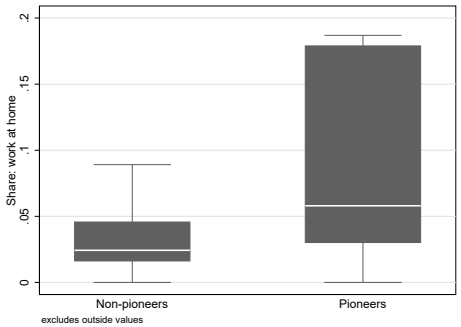
(e) have fewer children.



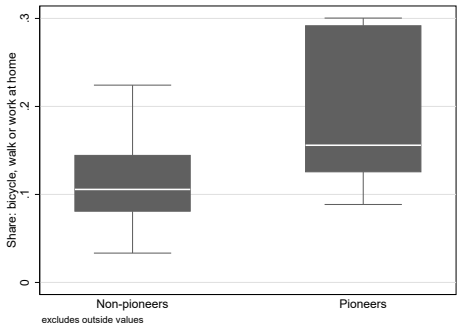
(f) are more often single.



(g) work more at home.



(h) commute shorter distances.



Notes: Our computations using IPUMS data for the New York metro area in 2000 and 2010. Following Costa and Kahn 2000, we define power couples as couples in which both members are college educated.

Table 6: Non-pioneer creative sectors, and sectors with similar worker characteristics

(a) Non-pioneer creative sectors		(b) Non-creative with similar characteristics	
451211	Book Stores	511110	Newspaper Publishers
451212	News Dealers and Newsstands	511140	Directory and Mailing List Publishers
545113	Prerecorded Tape, Compact Disc, and Record Stores	511199	All Other Publishers
512120	Motion Picture and Video Distribution	541511	Custom Computer Programming Services
512131	Motion Picture Theaters (except Drive-Ins)	541611	Administrative Mgmt and General Mgmt Consulting Services
512210	Record Production	541612	Human Resources and Executive Search Consulting Services
512230	Music Publishers	541614	Process, Physical Distribution, and Logistics Consulting Services
512290	Other Sound Recording Industries	541690	Other Scientific and Technical Consulting Services
517110	Internet Service Providers	541910	Marketing Research and Public Opinion Polling
519120	Libraries and Archives	541930	Translation and Interpretation Services
541320	Landscape Architectural Services	541940	Veterinary Services
541490	Other Specialized Design Services	541990	All Other Professional, Scientific, and Technical Services
541613	Marketing Consulting Services		
541712	R&D in the Physical, Engineering, and Life Sciences		
541720	R&D in the Social Sciences and Humanities		
541920	Photography Studios, Portrait		
611310	Colleges, Universities, and Professional Schools		
711110	Theater Companies and Dinner Theaters		
711310	Promoters of Performing Arts, Sports, and Similar Events with facilities		

Notes: Creative industries listed in column (a) are identified by the authors among sectors active in the New York MSA. Column (b) lists the non-creative, non-pioneer sectors whose workers' mean age is below the P75 of the mean age of pioneers' workers, whose share of educated workers is above the P25 of that for pioneer sectors, and whose share of workers working close to their home is above the P25 of that for pioneer sectors. For each industry, the table displays its 6-digit code and its label in the NAICS classification.

Table 7: Pioneers vs other creative sectors and non-creative sectors with similar workers.

	Coefficient on pioneers indicator / $[R^2]$					
Dependent variable	All sectors in NY		Creative sample		Similar sample	
Number of employees (log)	-0.481^a	[0.024]	-0.450^c	[0.062]	-0.263	[0.034]
Number of establishment (log)	0.497^c	[0.008]	1.274^a	[0.249]	0.038	[0.000]
Real wage (log)	0.117	[0.006]	-0.080	[0.019]	-0.283^a	[0.196]
Share of power couples	0.212^a	[0.078]	-0.027	[0.009]	-0.102^b	[0.127]
Share of college educated	0.216^a	[0.068]	-0.053	[0.037]	-0.093^c	[0.085]
Share working close to home	0.068^a	[0.088]	-0.011	[0.007]	0.008	[0.003]
Share working from home	0.048^a	[0.099]	-0.005	[0.002]	-0.011	[0.008]
Mean age	-1.888^a	[0.019]	0.854	[0.042]	-0.148	[0.001]
Share of married	-0.103^a	[0.072]	0.013	[0.010]	-0.036	[0.067]
Number of children	-0.292^a	[0.151]	0.034	[0.018]	-0.007	[0.001]
Sector included in the sample						
Pioneer sectors	Yes		Yes		Yes	
Non-pioneer creative sectors	Yes		Yes		No	
Non-creative but similar worker characteristics	Yes		No		Yes	
Other sectors considered in the paper	Yes		No		No	
Observations	422		45		38	

Notes: The table reports the coefficient estimated by a univariate regression where the left hand side variable is a sector characteristics and the explanatory variable is a dummy taking the value of one if the sector belongs to the list of pioneers. Coefficients are identified in the cross-section of sectors. Creative industries are identified by the authors among sectors active in the New York MSA. Non-creative, non-pioneer sectors are those whose workers' mean age is below the P75 of the mean age of pioneers' workers, whose share of educated workers is above the P25 of that of pioneer sectors, and whose share of workers working close to their home is above the P25 of that for pioneer sectors. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

Data Appendix

Census data. The area we consider as the NY metropolitan area comprises the following counties: Kings, Queens, New York, Suffolk, Bronx, Nassau, Westchester, Richmond, Orange, Rockland, Dutchess, and Putnam in the state of NY; and Bergen, Middlesex, Essex, Hudson, Monmouth, Ocean, Union, Passaic, Morris, Somerset, Sussex, and Hunterdon in the state of NJ. The area we consider as the Philadelphia metropolitan area comprises the following counties: Kent, and New Castle (DE); Cecil (MD); Atlantic, Burlington, Camden, Cape May, Cumberland, Gloucester, Salem in the state of NJ; and Berks, Bucks, Chester, Delaware, Montgomery, Philadelphia in the state of PA.

We extract block-level data for New York and for Philadelphia from the National Historical Geographic Information System (NHGIS) of the population center at the University of Minnesota (available at `https://www.nhgis.org`). All shape files we use are the 2010 Tiger versions provided by the Census. The algorithm described in online appendix O.6 provides us with a concordance that allows us to associate the 2010 blocks with stable units. We then use the shape files dissolved to the level of our harmonized blocks to assign the NETS establishments to blocks.

Census blocks are the most spatially disaggregated census units, with close to 250,000 blocks in the New York metropolitan area in 2010. We gather information on residents and housing units counts that are directly available at the block level. Several other variables—such as total income or the number of residents by educational attainment—are provided at a slightly higher level of aggregation, the block group. In that case, we apportion those variables to blocks using block-level population weights. Per capita and median household income, the age of the housing stock, as well as median rents and housing values are also available at the block-group level;

they are directly imputed to the blocks nested within the block groups.

National Establishment Time-Series. Walls & Associates teamed up with Dun and Bradstreet (D&B) to convert their national archival establishment data into a time-series database, the National Establishment Time-Series database (NETS). We use the 2014 version of that dataset for the New York core-based statistical area from 1990 to 2012, featuring more than 24 million geocoded establishment-year observations. Each establishment has a unique identifier (its DUNS number), latitude and longitude coordinates, a 6-digit NAICS industry code, and total employment at the establishment.

We use GIS software to assign each establishments in 1990, 2000, and 2010 to our 152,529 harmonized blocks, based on the latitude and longitude reported for each establishment in the data. We discard all establishments that are reported in the database but which do not fall into blocks of the New York metro area.

Other datasets. We supplement the census and the NETS data with several other datasets. These datasets are mainly used to create controls for our regressions.

We first obtain information on crime from the Furman Center for Real Estate and Urban Policy. Our crime data are reported at the precinct level and are, therefore, gathered only for the five boroughs of New York City. A few missing values are filled in with similar indicators from the *Historical New York City Crime Data* provided by the New York City Police Department. We use GIS software to map the crime data from the 75 precincts to the block level. For blocks that straddle several precincts, we compute the average number of crimes per capita across those precincts.

Turning to public transportation, the location of subway stations and of bus stops is provided by the Metropolitan Transportation Authority (MTA) and obtained from the NYC OpenData website. For metro stations located along the Metro-North and

Long Island Railroads, we use the publicly available *NYC Mass Transit Spatial Layers* produced by the GIS Lab at the Newman Library of Baruch College. Finally, the *New Jersey Geographic Information Network* provides us with similar information for lines operated by NJ TRANSIT as well as PATH (operated by Port Authority Trans Hudson) and PATCO (Port Authority Transit Corporation) lines. We then use GIS software to create a variable that gives the minimum distance of each block from a public transit stop. We also compute the number of lines in the neighborhood, and the number of stops (both bus and metro).

Regarding worker characteristics, we use IPUMS-USA data for the years 2000 (5% census data) and 2010 (5% ACS data) to compute NAICS-level indicators of worker characteristics by industry. We restrict our sample to employed workers and compute various characteristics at the 4-digit industry level (e.g., the share of college educated workers, of single workers, or of workers who commute by bicycle or by foot within each industry) for the entire U.S. and for the New York metropolitan area only.

We complement our dataset with geographic controls. First, we use the landmark datasets—both points and shapes—from the U.S. Census Bureau to create two variables. The first is a count of landmarks within each block as derived from the point pattern-based landmark files. The second is the minimum distance of each block to parks as contained in the shape-based landmark files. In the latter case, we keep only landmarks where the string 'Park' features in the name and drop all others (including those lacking a description). We further compute the distance of a block to the closest block composed exclusively of water and create a 'less than 200m from waterfront' dummy.

Turning to public housing and land-use information, we obtain information on the location of rent-controlled buildings in NYC and construct a variable that equals

the number of rent controlled units in the neighborhood. These data are publicly available from the github repository of the betaNYC project (https://beta.nyc and https://github. com/joepope44 /nyc_housing). The original data are provided by the NYC rent guidelines board (consult https://rentguidelinesboard. cityofnewyork.us/resources/rent- stabilized-building-lists/). The links between the tax assessment blocks and the census blocks is established using NYC's PLUTO shapefiles. The latter also provides information on vacant land, the presence of limited height districts, and the presence of historical districts. We construct indicators for whether a block belongs to a limited height districts, a historical districts; and we compute the share of vacant land in the neighborbood of each block.

Last, our data on the floor number of establishments comes from Liu et al. (2018) who use the extended version of the D&B dataset to retrieve the floor number of buildings from the establishments' address.[24] More specifically, they use the room or suite numbers to identify the floor. The data used to identify floors in our paper are for the universe of D&B establishments located in Manhattan in 2018. There are 448,759 establishments in that dataset. The floor number can be retrieved for 47.5% of these establishments. The latter are then assigned to the group of pioneers based on their main activity (NAICS 6-digit). As a robustness check, we have excluded from the sample all establishments located in 'skyscraper blocks'. We identify these blocks from the footprint outlines of buildings in New York City.[25] We define as 'skyscraper block' a block on which the average height of the buildings—weighted by the buildings' footprint—exceeds 300 feet. Results on this restricted sample are similar and available upon request.

[24]We thank Stuart Rosenthal for sharing the data on the floor number with us. More information on the data can be found in Liu et al. (2018).

[25]Data available at https://github.com/CityOfNewYork/nyc-geo-metadata/blob/master/ Metadata/Metadata_BuildingFootprints.md.

Data for Philadelphia. We also use data for the Philadelphia metropolitan area. The core of these data are the same as for New York: the census data from NHGIS and the NETS data from Wall & Associates. The data are processed in the same way as for New York. We do not have data on crime. Data on public transportation for Philadelphia come from the Southeastern Pennsylvania Transportation Authority (SEPTA) and are obtained from the *Pennsylvania Spatial Data Access* website. These data provide us with the location of regional rail, rapid transit rail, and trolley rail stations. As for New York, we compute the minimum distance of each block from a public transit stop using GIS software. Data on amenities (number of landmarks and distance to parks) are again constructed from the census landmark shapefiles. We finally construct a 'less than 200m from waterfront' dummy.

Gentrification and Pioneer Businesses: Supplemental Online Appendix

Kristian Behrens Brahim Boualam Julien Martin Florian Mayneris

O.1. Harmonized blocks

At what geographic scale should we analyze gentrification? Table O.3 in this online appendix shows that 21 out of 26 papers we reviewed work with the census tract geography, with five of them focusing solely on central city tracts. In our case, the effect of pioneer businesses on gentrification might be extremely local and tracts might be too large. We take a different approach and work at a finer geographic scale using time-consistent 'census blocks' for the metropolitan area at large. More precisely, we focus on time-consistent blocks in New York within a 30 kilometers radius around Wall Street (which we take as the city center, following Glaeser and Kahn, 2001).[26] These blocks represent around 60% of the population, establishments, and jobs in the New York metropolitan area over our study period.

Working at a fine geographic scale across a large portion of the metro area has two advantages. First, the block-level approach allows us to capture very localized dynamics that might wash out at a higher level of geographic aggregation such as tracts. We show indeed that pioneer businesses are more strongly associated with gentrification when the latter is measured more locally. Second, the broader view of the metro area allows us to look beyond the central city which is arguably special

[26] Baum-Snow and Hartley (2016) take a 4 kilometers radius around Wall Street, which captures about 2.8% of the metro population in our case. Couture and Handbury (2020) choose a variable distance cutoff to capture 5% of the metro population, which corresponds to a 5.5 kilometers radius around Wall Street.

in terms of population, income, and education dynamics. Although gentrification is often perceived as being a central-city phenomenon, we show that substantial gentrification occurs beyond the central city narrowly defined.[27]

The number and boundaries of census blocks—and of other census geographic units—change over time. In the greater New York metro area, the number of census blocks increased from 189,976 in 1990 to 240,318 in 2010. This increase masks a wide range of changes made by the census to the geography of these blocks. Some are split while others are grouped together. More problematic, some blocks are split and their parts recombined in complex ways into several new or existing blocks. Since blocks are defined based on population counts, these problems affect especially areas with strong population dynamics that may be of interest for the analysis of gentrification. To deal with these problems, we develop an algorithm that can be used to create constant geographies based on census blocks (see online Appendix O.6 for details). We refer to those blocks as *harmonized blocks* (or blocks, for short).

Table O.6 in this online appendix reports the distribution of the average number of census blocks per harmonized block in New York for 1990, 2000, and 2010. More than 75% of our harmonized blocks consist of a single census block, more than 90% are made of 1 or 2 census blocks, and more than 95% are made of 1 to 3 census blocks. Less than 1% of our harmonized blocks contain more than 8 census blocks. On average, a harmonized block contains 1.4 census blocks, which means it is much smaller than either a census block-group (which has, on average, 16 census blocks)

[27]There is a third purely technical advantage. As discussed in the next section, we have to make the geographic units time consistent. Since we lack official crosswalks at the level of block-groups (smaller than tracts but bigger than blocks, and at the level of which average income per capita and population counts by education level are available in census data), harmonizing them produces time-consistent units that are far larger than the ones we will use, thus negating the benefits of a finer geographic scale.

or a census tract (which has, on average, slightly more than 50 census blocks).

O.2. Descriptive statistics.

Table O.1 shows the characteristics of harmonized blocks in terms of population size, per capita income, and the share of educated residents (see online appendix O.6 for additional details on the construction of these blocks; and Table O.7 for additional descriptives). Harmonized blocks have on average about 185 residents. The population distribution across blocks is skewed since the median number of residents is only about half of the average.[28] This contrasts with the distributions of per capita income and the share of educated—defined as those with at least some college degree—where the median, though lower than the average, is not far from the latter.

Table O.1: Characteristics of harmonized blocks, 1990–2010.

	1990, Percentile				2000, Percentile				2010, Percentile			
	Mean	25	50	75	Mean	25	50	75	Mean	25	50	75
# residents	187.4	51	98	207	182	40	92	210	185.7	41	94	217
Per capita income	18,763	12,692	16,740	20,849	25,840	16,125	22,376	29,323	33,721	20,971	29,022	39,119
Share educated	0.2	0.10	0.17	0.26	0.23	0.12	0.20	0.31	0.27	0.15	0.25	0.37

Notes: Average characteristics of all harmonized blocks whose centroids are less than 30 kilometers from Wall Street and which are not exclusively composed of water. There are 63,799 such harmonized blocks in total.

O.3. Changes in income and education.

This appendix shows descriptive evidence on income and education changes at the block level. We show that the dynamics of income and education are not perfectly

[28] This skewness is even more striking for the number of establishments and jobs, in line with the well-documented fact that economic activity generally displays more geographic concentration than population. See Table O.7 in this online appendix.

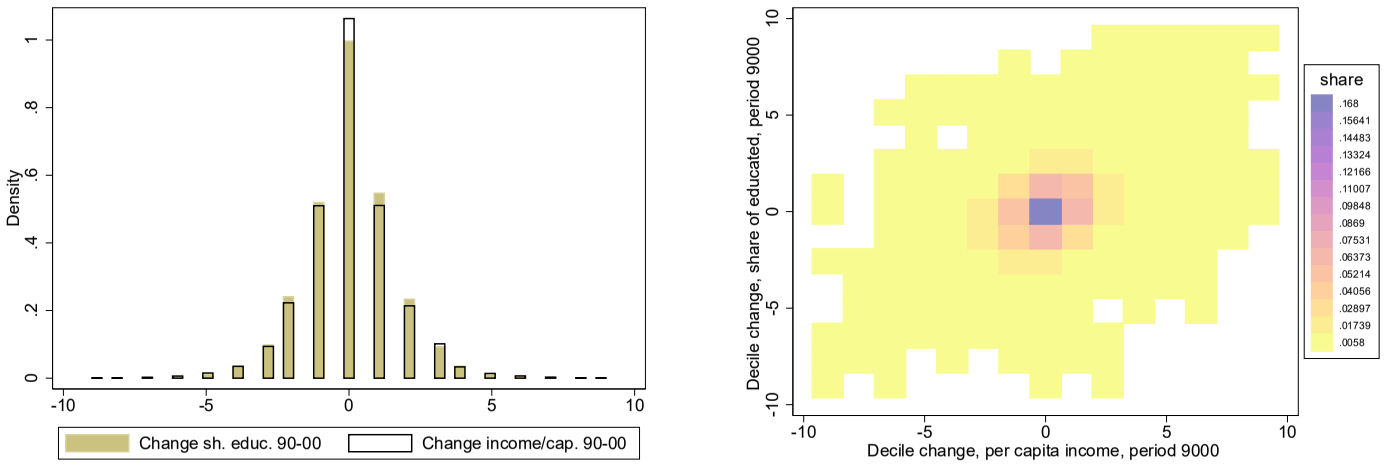
spatially correlated—especially for initially poorer blocks—which justifies the construction of our three distinct measures of gentrification. We further show that there is a lot of idiosyncrasy in income and education changes across narrowly defined blocks within tracts, thus justifying the granular spatial scale at which we work.

Block-level changes. As shown in panel (a) of Figure O.1, the patterns of education or income mobility—whether depicted using histograms or heatmaps—look quite similar when considering all blocks. Although there is slightly less mobility in income than in education, both distributions are fairly symmetric, with most blocks not moving much and a few blocks transitioning quickly up or down. If one focuses on poor blocks only, as in panel (b) of Figure O.1 and defined as blocks with initial income below the metropolitan median, we see a more skewed upward mobility in terms of income and a more diffuse mobility in terms of education. In particular, there are many blocks with substantial upward income mobility but little upward educational mobility. This contrast explains why our discrete definition M_1 of gentrification puts more stringent conditions on changes in income than changes in education.

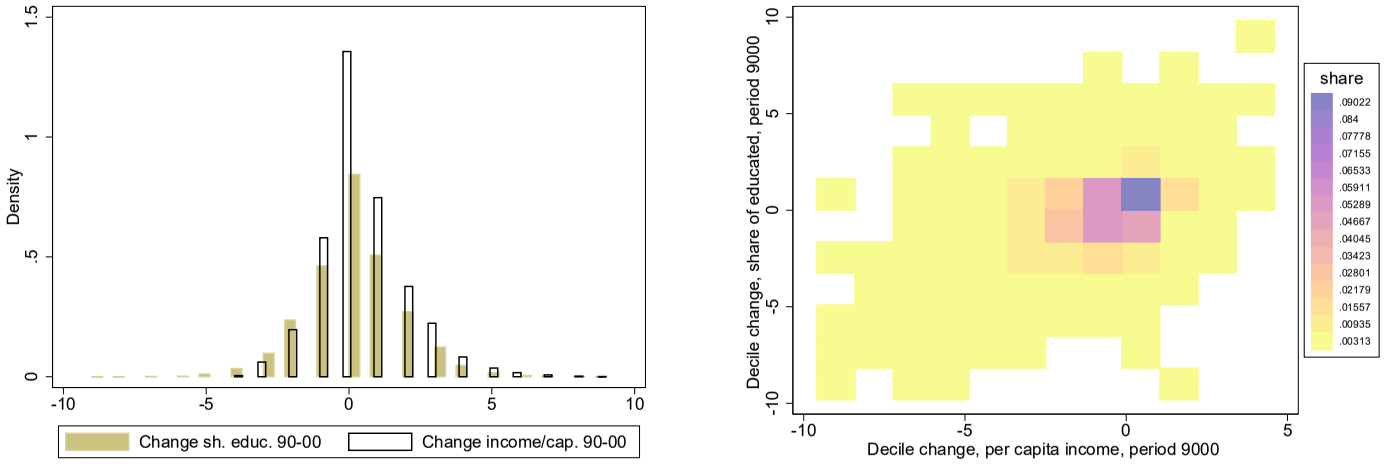
Between tract variation. Table O.2 reports the contribution of the between-tract variation in levels and changes in per capita income and share of educated to overall variation of these variables. It shows substantial heterogeneity within census tracts in terms of income and education. This is especially obvious for changes as shown by the low R^2 of the regressions of block-level income and share of educated changes on tract fixed effects, which explain less than half of the observed variation. This suggests that there is a lot of idiosyncrasy in income and education changes across narrowly defined blocks within tracts. This finding suggests that using fine-grained data at the block-level may help us improve the detection of gentrification hotspots

Figure O.1: Mobility of blocks by per-capita income and share of educated residents (1990–2000).

(a) all harmonized blocks



(b) blocks with income below the median



Notes: Distribution of the change in income and education decile of harmonized blocks between 1990 and 2000.

Table O.2: Contribution of between-tract variations to overall block-level variations in income and education.

	Years	(1) FE contribution	(2) FE reg. R^2	(3) RE contribution
Per capita income	1990	0.87	0.85	0.86
	2000	0.87	0.89	0.89
	2010	0.87	0.84	0.87
Share of highly educated	1990	0.87	0.84	0.87
	2000	0.89	0.87	0.89
	2010	0.79	0.75	0.79
Change, ln per capita income	1990–2000	0.54	0.20	0.51
	2000–2010	0.85	0.44	0.85
Change, share of educated	1990–2000	0.56	0.39	0.53
	2000–2010	0.45	0.30	0.41

Notes: This table shows the contribution of the between-tract variation to the overall variation observed across blocks in levels and changes for income and the share of educated. We compute the contribution of between-tract dispersion to overall dispersion as $var(u)/[var(u) + var(e)]$ in columns (1) and (3), where u are the tract fixed effects (FE) or random effects (RE) and e is the error term. We also report in column (2) the R^2 obtained from regressions of each of the four variables on tract fixed effects.

and their connection with the presence of pioneer businesses.

O.4. Bibliometric analysis.

Table O.3 summarizes key dimensions of 27 widely cited and/or recent papers in economics, urban planning, and sociology that we reviewed in search of a definition for gentrification.

Table O.3: Definitions of gentrification used in the literature.

	Field	GS citations	Definition	Geography	Eligibility criteria	Income	Housing price	Share renters	Share edu.	Age	Race	Multi
Barton (2016)	urban	75	yes	hood	yes	yes	yes	yes	yes	yes	yes	yes
Baum-Snow & Hartley (2016)	econ	29	no	tract	dwntwn	no	no	no	yes	no	no	no
Bostic and Martin (2003)	urban	129	yes	tract	yes	yes	no	no	no	no	no	no
Bostic and Martin (2003) (measure 2)	urban	129	yes	tract	yes	yes	yes	yes	yes	yes	yes	yes
Brueckner and Rosenthal (2009)	econ	356	no	tract	dwntwn	yes	no	no	no	no	no	no
Brummet and Reed (2018)	econ	2	yes	tract	yes	no	no	no	yes	no	no	no
Couture et al. (2018)	econ	36	yes	tract	dwntwn	yes	no	no	no	no	no	no
Couture and Handbury (2018)	econ	73	no	tract	dwntwn	no	no	no	yes	no	no	no
Ding et al. (2016)	econ	136	yes	tract	yes	no	yes	no	yes	bo	no	yes
Edlund et al. (2015)	econ	82	no	tract	no	no	yes	no	no	no	no	no
Ellen and O'Regan (2011)	econ	175	yes	tract	yes	yes	no	no	no	no	no	no
Ellen et al. (2017)	econ	37	yes	tract	yes	yes	no	no	yes	no	no	yes
Freeman (2005)	urban	647	yes	tract	yes	no	yes	no	yes	no	no	yes
Freeman and Braconi (2004)	urban	712	yes	district	yes	no	no	no	no	no	no	no
Glaeser et al. (2018)	econ	38	no	zipcode	no	no	yes	no	no	no	no	no
Grodach et al. (2014)	urban	114	no	zipcode	no	yes	yes	yes	yes	yes	yes	yes
Guerrieri et al. (2013)	econ	358	yes	tract	yes	no	yes	no	yes	no	no	yes
Hammel and Wylly (1996)	urban	162	yes	tract	yes	yes	yes	yes	yes	yes	yes	yes
Hwang and Lin (2016)	socio	50	yes	tract	no	yes	no	no	yes	no	no	yes
Lee (2003)	urban	665	yes	case std.	yes	yes	no	no	no	no	no	no
Lester and Hartley (2014)	econ	34	yes	tract	yes	no	yes	no	yes	no	no	yes
McKinnish et al. (2010)	econ	302	yes	tract	yes	yes	no	no	no	no	no	no
Meltzer (2010)	urban	30	yes	tract	yes	yes	no	no	no	no	no	no
Meltzer and Ghorbani (2017)	econ	30	yes	tract	yes	yes	no	no	no	no	no	no
O'Sullivan (2005)	econ	51	yes	tract	yes	yes	no	no	no	no	no	no
Su (2018)	econ	17	no	tract	dwntwn	no	no	no	yes	no	no	no
Zukin et al. (2009)	socio	481	no	hood	yes	no	no	no	no	no	no	no

Notes: field is the field in which the paper is published (economics, urban affairs/planning, sociology); *GS citations* is the number of Google Scholar citations as of July 2020; *definition* is 'yes' if the authors provide a definition of gentrification; *geography* is the level at which gentrification is measured (census tract, district, zipcode, neighborhood, or specific case study); *eligibility* indicates whether the study uses an eligibility criterion—it is 'yes' if gentrification applies only to poor neighborhoods ('dwntwn' (downtown) means the eligibility is about location rather than income); *income*, *housing prices*, *share renters*, *share edu.*, *age*, and *race* indicate whether the definition uses information on income, housing prices, the share of renters, the share of educated, the age or residents, and the racial composition of residents. Finally, *multi* indicates whether more than one criterion is used in the definition.

O.5. Additional information on NETS data.

One important feature of the NETS data for our purpose is the location information of the establishments.[29] Depending on the precision of the geocoding, the latitude and the longitude reported in the NETS data are mainly based on either 'rooftop' or ZIP-code. 'Rooftop' means that all the criteria for an exact address have been met. "ZIP-code" means that the exact address could not be determined and that the centroid of the corresponding ZIP-code is used as an approximate location (which can be more precise for establishments than, e.g., census tracts).[30] Panel (a) of Table O.4 summarizes the accuracy of the geocoding in our dataset. It shows that three-quarter of the establishments, accounting for 77% of employment, are rooftop geocoded in 1990. The corresponding figures increase over time and stand at 96.6% and 94.4% in 2010, respectively.

Turning to the number of establishments and their size distribution, panel (b) of Table O.4 shows that the total number of establishments reported in the NETS data almost doubled in 20 years. It increases from about 650,000 in 1990 to about 1.3 millions in 2010. This feature is driven both by an increasing coverage of the D&B data and by a large increase in SIC 73899999 ('Business activities at non commercial sites', according to the D&B classification). The latter industry displays an abnormally large increase in the number of its establishments—going from about 900 in the early 2000 to 115,000 in the early 2010. It includes all types of electronic micro businesses, such as private persons who sell items through electronic platforms such

[29] See Walls and Associates (2014) and Neumark et al. (2011) for more information on NETS data.

[30] D&B underline that ZIP-codes may allow for more accurate positioning of businesses than census tracts or ZIP-code tabulation areas (ZCTA) of the Census Bureau. Although there are fewer ZIP codes than census tracts, ZIP codes may in many instances be more accurate for businesses than the alternative census geographies as many large office buildings or industrial complexes can have their own ZIP code.

Table O.4: Geocoding and sectoral breakdown of the NETS data for New York.

(a) Accuracy of geocoding						
	Share of establishments			Share of employment		
Geocoding type	1990	2000	2010	1990	2000	2010
Block face	73.4%	85.9%	96.6%	77.4%	87.7%	94.4%
ZIP-code	25.5%.	12.8%	2.1%	20.3%	9.5%	2.9%
Others	1.1%	1.3%	1.3%	2.3%	2.8%	2.7%
(b) Establishment size distribution						
# of employees	1990	2000	2010			
1	108,735	200,569	376,629			
2 to 5	329,214	439,527	671,104			
6 to 10	96,368	103,409	97,080			
11 to 50	95,810	105,606	99,927			
50+	25,869	27,524	26,981			
Total	655,996	876,635	1,271,721			

Notes: Panel (a) reports the share of establishments and employment in New York by accuracy of their geocoding in the NETS data. Panel (b) reports the number of establishments by size category as well as the total number. All figures are for the NETS New York CBSA dataset.

as eBay or Etsy and have registered a business at home for doing so. Since this sector does not stand out as being particularly important for gentrification in our analysis—it is not a pioneer sector—this large increase should not be an issue.

One may wonder how the NETS data compare with other U.S. establishment-level data. It is worth noting that NETS data, census data, and Bureau of Labor Statistics (BLS) data do not cover the same establishments. Indeed, the NETS cover the self-employed while the other two datasets do not. Furthermore, the definition of an establishment differs across datasets. In the NETS data, an establishment is defined as a unique location and a unique primary market. This explains why the NETS data report on average 2.5 times more establishments in 2012 than the County Business Patterns in the five boroughs of New York (Bronx, Kings, New York, Queens, and Richmond counties).

O.6. Geographic concordance algorithm.

We provide details on the algorithm we use to harmonize census blocks over time. We start with a simple example to explain our graph-theoretic approach to building concordances. Table O.5 describes the structure of correspondence for a hypothetical nomenclature revised between years 1 and 2, and then again between years 2 and 3. For instance, in observations [1] and [2], code a is split into codes a and b between years 1 and 2. Also, as can be seen from observation [3], the name of code d is modified between years 1 and 2. Between years 2 and 3, summarized in the latter half of Table O.5, both codes a and b are split into codes b and c. Furthermore, code e is split into codes a and d, the latter one being recycled after having been retired between years 1 and 2.

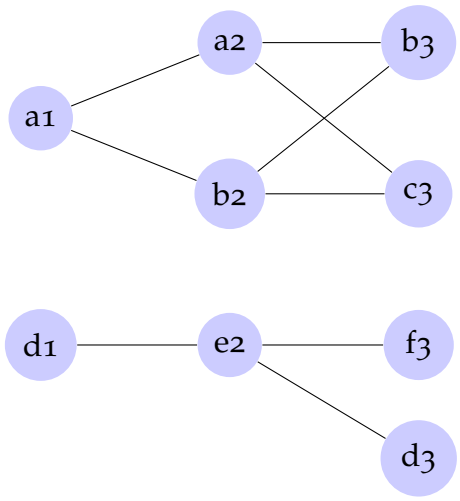
Table O.5: Sample correspondence table.

		Old			New		
Years	Obs	Code	Partial flag	Year	Code	Partial flag	Year
1-2	[1]	a	p	1	a		2
1-2	[2]	a	p	1	b		2
1-2	[3]	d		1	e		2
2-3	[4]	b	p	2	b	p	3
2-3	[5]	b	p	2	c	p	3
2-3	[6]	a	p	2	b	p	3
2-3	[7]	a	p	2	c	p	3
2-3	[8]	e	p	2	f		3
2-3	[9]	e	p	2	d		3

Notes: Example with three years. Statistical agencies would provide one table for the passage from year 1 to 2 (top panel), and a separate table for the passage from year 2 to 3 (bottom panel). 'Partial flag' identifies modifications that are not 1 to 1.

Observe that the correspondence in Table O.5 has the same structure as the correspondence tables generally provided by statistical agencies (e.g., it is similar to that

Figure O.2: Example of connected components.



used by the Census Bureau in its geographical relationship files). It may be viewed as describing a *correspondence graph*, where the combination code-year uniquely identifies a node and where the correspondence relationships are the edges. Being a graph, the correspondence in Table O.5 induces an adjacency matrix. It contains all the 'ones', but not the 'zeros'. The zeros are all possible combinations of the nodes (the codes) which are not directly linked.

Figure O.2 displays the graph associated with Table O.5. Each node (e.g., a1 or d3) corresponds to a unique code-year combination. As can be seen, optimally harmonizing the codes of Table O.5 requires finding the smallest groups of codes that are all connected and thus define components that are invariant and comparable over time. Figure O.2 shows that there are two connected components in our graph. This means that we can build two synthetic groups of related codes: $G_1 = \{\text{a1}, \text{a2}, \text{b2}, \text{b3}, \text{c3}\}$ and $G_2 = \{\text{d1}, \text{d3}, \text{e2}, \text{f3}\}$. The time-invariant smallest synthetic groups (SSG) of codes are the connected components of the graph whose nodes are the codes and whose edges are given by the revisions of the nomenclature (i.e., the relationship files). Any concordance problem based on crosswalks provided by statistical agencies can be

viewed as in Table O.5 and Figure O.2. Hence, we can approach concordance problems in very general terms and propose a method that is applicable to all of them. Our algorithm—in pseudo code—is as follows:

Algorithme 1 : Connected components concordance (C^3)

Data : *In a preliminary step*, build a 3-columns file with old and new codes variables (given by unique code-year identifiers) and an edge variable set to one. The file is saved in `ascii`.

Result : Codes and their synthetic groups saved in the `ascii` file `corres.txt`.

1: Load the data in `Matlab`
2: Build the adjacency matrix
3: Identify the connected components (using `networkComponents.m`)
4: Assign a unique identifier to each connected component (these unique numbers identify the synthetic groups that constitute the concordance)
5: Save the data in an `ascii` file

This algorithm builds on the observation that the optimal concordance (i.e., the SSG) corresponds simply to finding the connected components of the graph spanned by the code-year nodes and the revision edges. Once viewed in these terms, it becomes a relatively standard problem that can be solved efficiently using the tools of graph theory to find the connected components and to build synthetic identifiers for related codes.[31] This method is simple, extremely efficient, universally applicable, produces minimum concordances, and can be readily implemented using standard

[31]Once the problem is viewed in these terms, it becomes clear that all concordance problems can be approached in exactly the same way. Making use of standard tools from graph theory, large problems involving many years and hundreds of thousands of units can be solved very efficiently. Previous methods on census blocks create 'standardized blocks' between consecutive census years, and then iterate across years (see Carillo and Rothbaum 2016 for an application to Washington DC), whereas our method deals with all years simultaneously.

software packages. It is also not affected by a number of problems that plague more specific algorithms (for example, recycling retired identifiers over time poses no problem for our method).[32] We use `Stata` to prepare the intermediate data, and `networkComponents.m`, an open source `Matlab` code by Daniel Larremore, to find the connected components of the graph.

O.7. Descriptives of harmonized blocks

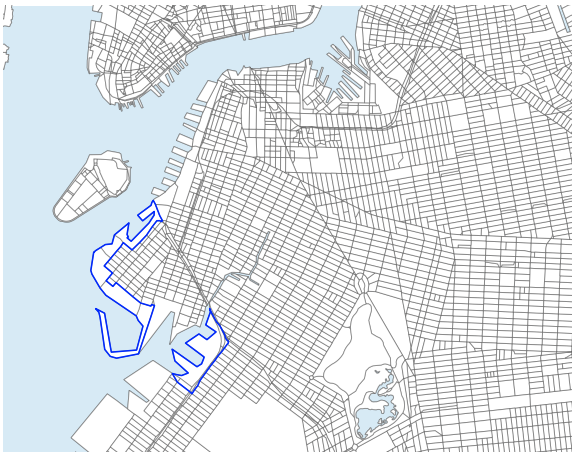
As is well known the smallest synthetic groups (SSGs) are, by definition, more aggregated than the original units from which they are constructed. This naturally raises the question of how much geographic information we lose when harmonizing census blocks over time. Our algorithm identifies 152,529 time-consistent *harmonized blocks* (henceforth blocks, for short) for New York for the period 1990–2010. This set of blocks corresponds to the smallest stable census geography for these two decades. Dropping all blocks that have either zero population or that consist only of water leaves us with 150,747 blocks, which is our time-consistent geography in the subsequent analysis.

Panel (a) of Figure O.3 provides an illustration of how our harmonized blocks (in blue) relate to census blocks (in gray). Table O.6 summarizes these relationships for the whole New York metro area. As shown, more than 75% of our harmonized blocks consist of a single census block, more than 90% are made of 1 or 2 census blocks,

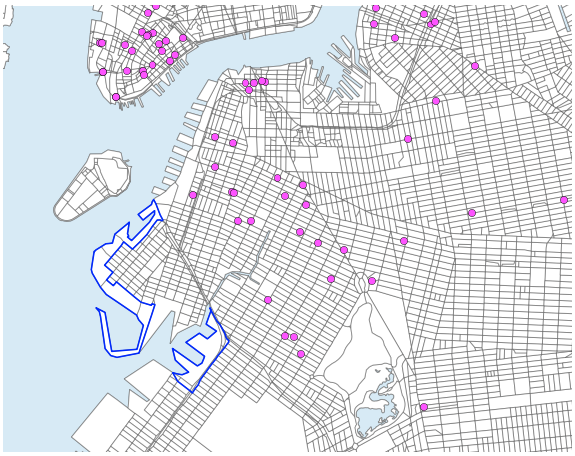
[32] There are, e.g., several papers that concord product nomenclatures over time (see Pierce and Schott 2012 for U.S. product categories; Martin and Mejean 2014 for the French product nomenclature; and Bernard et al. 2012 for EU product categories and industries). It is fair to say that those approaches are usually custom-tailored to specific product datasets and all rely on adaptations of the algorithm developed by Pierce and Schott 2012. They are hence not portable. Tests that we ran also suggest that they are much slower than our approach for large datasets.

Figure O.3: Harmonized blocks versus census blocks and granularity of the NETS data.

(a) Illustration of harmonized blocks.



(b) Granularity of NETS data.



Notes: Harmonized blocks, 1990–2010, as determined by the methodology explained in online appendix O.6. Panel (a) shows the relationship between census blocks (in grey) and selected harmonized blocks (in blue). Panel (b) depicts the location of art dealers (NAICS 453920) in northern Brooklyn and southern Manhattan in 2000.

and more than 95% are made of 1 to 3 census blocks only. Less than 1% of our harmonized blocks contain more than 8 census blocks.[33] On average, a harmonized block contains 1.4 census blocks, which means that it is much smaller than either a census block-group (which has, on average, 16 census blocks) or a census tract (which has, on average, slightly more than 50 census blocks). As expected, the size of our harmonized blocks increase as we move away from the central city. The reason is that census geography revisions are more frequent in places that experience substantial

[33] Although our blocks are fairly small on average, a few of them contain more than 50 to 100 census blocks. Those large synthetic blocks are mainly located in the outskirts of the metropolitan area—more than 30 kilometers from Wall Street—where more rapid urban expansion leads to a substantial redefinition of census blocks. However, there are also some more central areas that are prone to successive redefinitions between census years and that are important for our analysis: waterfronts, parks, and urban redevelopments in formerly non-residential areas. Since anecdotal evidence suggests that many developments along the waterfront correspond to gentrification, it is especially important to have a constant geography there to precisely capture demographic and socio-economic changes over time.

population changes, which are generally newly constructed zones in the outlying zones of the metro area.

Table O.6: Number of census blocks per harmonized block, 1990–2010.

	Percentile					Max	Mean
	50	75	90	95	99		
# of census blocks, all of New York metro area	1	1	2	3	7.7	157.7	1.4
# of census blocks, 30 km around Wall Street	1	1	1.7	2	4.3	74.7	1.2

Notes: This table reports the distribution of the average number of census blocks per harmonized block in New York for 1990, 2000, and 2010. There is a total of 152,529 harmonized blocks in our dataset. A number of census blocks equal to 1 in the table means that, on average over 1990–2010, the harmonized block consists of a single census blocks (i.e., it is stable). The first line provides the information for all census blocks in the New York metro area, whereas the second line provides the information for the part of the city within a 30 kilometers radius around Wall Street. In the latter case, we have 63,799 blocks.

O.8. Problems when working at the block-group level.

Concordance issues. Since some of our census variables are only reported at the block-group level, one may ask why we do not directly work at that level (which would allow us to sidestep the question of apportioning the variables from block-groups to blocks). Our concordance algorithm can be applied to other spatial units as well, so we have experimented with block groups. We ran into an important problems. As for blocks, the boundaries of block groups change between census years. Yet, contrary to blocks, there are no relationship files linking block groups over time. To nevertheless get an idea, we ran our concordance algorithm on block groups, using the block relationship files and aggregating them up to the block-group level. The resulting *harmonized block groups* contain on average 1.8 census block groups, with 10% having more than 3. The concordance is especially problematic

in some areas, including the waterfront. For example, a single harmonized block group gathers all block groups bordering both the Hudson river and the East river in Manhattan. Similar problems arise for Staten Island. Hence, the time-consistent geography for census block groups obtained using our algorithm is much coarser than that obtained for blocks. This coarse geography makes the identification of highly localized dynamics extremely difficult.

Identification of pioneer sectors. We also used a *subset* of stable block groups (defined as those who did not change across time) and ran our identification of pioneers at that level (given the size of block groups, we did not aggregate the variables to the neighborhood using 250m, 500m radii or contiguity and directly worked at the block group). As Table 3 in the main text shows, we identify fewer pioneer sectors but all those that we do identify belong to our base list of sectors.

O.9. Additional tables and figures

Table O.7 summarizes other characteristics than population, income, and education at the level of our harmonized blocks. Tables O.8 and O.9 show the detailed results for our baseline estimations for New York and Philadelphia. Table O.10 reports estimations where we interact changes in pioneers with distance to high-skilled employment centers and natural amenities. Table O.11 shows that industries that are creative but not pioneers, and industries with similar worker characteristics as pioneers but which are not pioneers, do not correlate positively with subsequent gentrification. Finally, Figure O.4 shows the distribution of selected worker characteristics for workers in pioneer sectors and workers in non-pioneer sectors in the U.S.

Table O.7: Other characteristics of harmonized blocks, 1990–2010.

	1990, Percentile				2000, Percentile				2010, Percentile			
	Mean	25	50	75	Mean	25	50	75	Mean	25	50	75
# residents	187.4	51	98	207	182	40	92	210	185.7	41	94	217
Per capita income	18,763	12,692	16,740	20,849	25,840	16,125	22,376	29,323	33,721	20,971	29,022	39,119
Share educated	0.2	0.10	0.17	0.26	0.23	0.12	0.20	0.31	0.27	0.15	0.25	0.37
Median gross rent	643.6	519	625	746	868.4	702	798	946	1,239.5	998	1,176	1,426
Median housing value	203.1	155.7	186	225.4	237.9	164.6	199.6	263	497.2	373.4	455	595.6
# establishments	6	0	1	4	7.8	0	2	6	11.3	2	5	11
# jobs	95.6	0	3	28	100.3	0	7	41	95.3	3	12	50

Notes: This table presents the average characteristics of all harmonized blocks whose centroids are less than 30 kilometers from Wall Street and which are not exclusively composed of water. There are 63,374 such harmonized blocks in total. All monetary values are expressed in current U.S. dollars, except median housing values which are expressed in thousands.

Table O.8: Pioneers and gentrification in New York (500m), 2000–2010.

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
Δ # pioneer estab.	1.389^a	1.009^a	1.778^a	1.498^a	0.399^a	0.266^a
	(0.311)	(0.244)	(0.323)	(0.273)	(0.124)	(0.090)
Δ Ln (1+ # non pioneer estab)	0.008	-0.002	0.012	0.001	0.011^b	-0.000
	(0.010)	(0.013)	(0.013)	(0.015)	(0.004)	(0.006)
Ln per cap. income	-0.019	0.024	-0.269^a	-0.195^a	0.058^a	0.077^a
	(0.026)	(0.034)	(0.040)	(0.049)	(0.010)	(0.014)
Share college edu. resid.	0.533^a	0.382^a	1.096^a	0.790^a	0.023	-0.059
	(0.103)	(0.135)	(0.112)	(0.139)	(0.042)	(0.053)
Ln rent	-0.022	-0.010	-0.011	0.002	-0.013^b	-0.007
	(0.014)	(0.018)	(0.017)	(0.022)	(0.005)	(0.007)
Median age of buildings	0.002^a	0.002^a	0.001^c	0.002^b	0.000^b	0.001^a
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Share black resid.	-0.009	0.001	-0.036^c	0.016	0.009	0.023^b
	(0.016)	(0.023)	(0.019)	(0.029)	(0.006)	(0.010)
Share asian resid.	-0.192^a	-0.201^a	-0.314^a	-0.271^a	-0.032^c	-0.043^b
	(0.046)	(0.049)	(0.051)	(0.055)	(0.018)	(0.020)
Share other resid.	-0.004	0.003	-0.188^a	0.009	0.025	0.055^b
	(0.051)	(0.060)	(0.065)	(0.068)	(0.020)	(0.024)
Ln pop.	-0.009	-0.036^a	-0.008	-0.057^a	0.004	-0.016^a
	(0.006)	(0.012)	(0.009)	(0.016)	(0.003)	(0.005)
Less than 200m from waterfront	0.013	0.019	0.013	0.040^b	0.003	0.003
	(0.013)	(0.023)	(0.014)	(0.020)	(0.005)	(0.007)
Ln (1+# train lines)	0.014^b	0.002	0.017^a	0.004	0.008^a	0.002
	(0.006)	(0.006)	(0.007)	(0.007)	(0.002)	(0.002)
Ln (1+# bus lines)	-0.003	-0.012	-0.007	-0.000	-0.001	-0.003
	(0.005)	(0.011)	(0.005)	(0.010)	(0.002)	(0.003)
Distance to closest park (log)	–0.004^c	-0.007^b	-0.001	-0.002	-0.001	-0.001
	(0.002)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)
Ln # of main landmarks	-0.001	0.016^c	0.006	0.017^c	0.004	0.005
	(0.006)	(0.009)	(0.008)	(0.010)	(0.003)	(0.003)
Socio-economic changes in the neighborhood 1990–2000	-0.009^b	-0.004	0.120^a	0.069	-0.078	0.023
	(0.004)	(0.005)	(0.043)	(0.052)	(0.065)	(0.085)
# murder per cap.		-0.389^b		-0.557^a		-0.116^c
		(0.156)		(0.201)		(0.066)
# burglary per cap.		-0.022^a		-0.031^a		-0.012^a
		(0.006)		(0.008)		(0.003)
# robbery per cap.		0.031^a		0.046^a		0.013^a
		(0.006)		(0.007)		(0.002)
# rape per cap.		-0.081		-0.149		-0.059
		(0.112)		(0.131)		(0.042)
Ln (1+# rent control buildings)		0.014^a		0.010^b		0.008^a
		(0.005)		(0.005)		(0.002)
Share vacant land		0.077		0.032		-0.016
		(0.161)		(0.219)		(0.048)
	0.051		-0.140^c		0.012	
		(0.253)		(0.083)		(0.045)
Presence of historical districts		0.012		0.026		0.008
		(0.015)		(0.017)		(0.006)
# of observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.047	0.072	0.055	0.078	0.035	0.072
Sample	New York	NYC	New York	NYC	New York	NYC

Notes: Reported coefficients and standard errors are mutiplied by 1,000 compared to the actual ones for variable "Δ # pioneer estab.". The sample is composed of blocks with per capita income below the median in the city in 2000, and with at least eight residents. The measure of exposure to pioneers is given by equation (3) in the main text. All explanatory variables are measured in 2000 and are computed using 500 meters rings around each block (except the distance to subway, to parks, and to closest gentrifying block variables, as well as the waterfront dummy). Robust standard errors, corrected for cross-sectional spatial dependence within a 500 meters radius (using HAC estimation), are reported in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%. The proxy for "Socio-economic changes in the neighborhood 1990–2000" is the distance to the closest block that gentrified between 1990 and 2000 when our gentrification indicator is used as a dependant variable, and the change in average per capita income or in the share of educated residents between 1990 and 2000 for our two other proxies for socio-economic changes.

Table O.9: Determinants of gentrification in Philadelphia (500m), 2000–2010.

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
Δ # pioneer estab.	18.183^{a}	31.757^{a}	6.988^{a}	15.756^{c}	1.574^{a}	4.139^{b}
	(2.938)	(12.073)	(2.313)	(8.472)	(0.443)	(1.676)
Δ Ln (1+ # non pioneer estab)	-0.048	-0.064	0.035	0.024	-0.001	-0.005
	(0.038)	(0.047)	(0.023)	(0.031)	(0.006)	(0.008)
Ln per cap. income	0.044	0.068	-0.419^{a}	-0.400^{a}	0.039^{a}	0.041^{a}
	(0.115)	(0.116)	(0.068)	(0.070)	(0.013)	(0.014)
Share college edu. resid.	0.469	0.173	1.252^{a}	1.048^{a}	-0.137^{a}	-0.188^{a}
	(0.306)	(0.368)	(0.170)	(0.212)	(0.042)	(0.047)
Ln rent	-0.054^{c}	-0.044	-0.033	-0.028	-0.013^{a}	-0.011^{b}
	(0.033)	(0.034)	(0.026)	(0.025)	(0.005)	(0.005)
Median age of buildings	0.004^{a}	0.003^{c}	0.001	0.000	0.000^{a}	0.000
	(0.001)	(0.002)	(0.001)	(0.001)	(0.000)	(0.000)
Share black resid.	-0.113	-0.092	-0.140^{a}	-0.124^{a}	-0.019^{c}	-0.016
	(0.069)	(0.072)	(0.034)	(0.038)	(0.010)	(0.011)
Share asian resid.	0.376	-0.687^{c}	0.004	-0.216	-0.014	-0.077
	(0.316)	(0.392)	(0.235)	(0.242)	(0.037)	(0.049)
Share other resid.	-0.371^{c}	-0.380^{b}	-0.353^{a}	-0.349^{a}	-0.035	-0.037
	(0.190)	(0.189)	(0.116)	(0.115)	(0.025)	(0.026)
Ln pop.	0.031	0.009	-0.053^{a}	-0.070^{a}	-0.007^{b}	-0.011^{a}
	(0.021)	(0.028)	(0.015)	(0.021)	(0.003)	(0.004)
Less than 200m from waterfront	0.016	0.023	-0.010	-0.007	-0.002	-0.001
	(0.031)	(0.032)	(0.016)	(0.016)	(0.004)	(0.004)
Ln Distance to subway	-0.006	-0.006	0.000	0.001	-0.002^{c}	-0.002^{c}
	(0.010)	(0.010)	(0.005)	(0.005)	(0.001)	(0.001)
Ln Distance to closest park	0.014	0.012	0.009^{b}	0.007^{c}	0.001	0.001
	(0.010)	(0.010)	(0.004)	(0.004)	(0.001)	(0.001)
Ln (1+# of main landmarks)	-0.047^{c}	-0.053^{b}	0.000	-0.004	0.001	-0.001
	(0.024)	(0.025)	(0.013)	(0.013)	(0.003)	(0.003)
Socio-economic changes in the neighborhood 1990–2000	-0.053^{a}	-0.041^{b}	0.054	0.017	-0.093^{c}	-0.113^{b}
	(0.016)	(0.018)	(0.056)	(0.065)	(0.051)	(0.054)
# of observations	18,144	18,144	18,009	18,009	18,141	18,141
R-squared	0.157	n.a.	0.064	n.a.	0.088	n.a.
Specification	LPM	IV	OLS	IV	OLS	IV
Kleinbergen-Paap F-stat	n.a.	7.925	n.a.	7.929	n.a.	8.028

Notes: Reported coefficients and standard errors are mutiplied by 1,000 compared to the actual ones for variable "Δ # pioneer estab.". The sample is composed of blocks with per capita income below the median in the city in 2000, and with at least eight residents. The measure of exposure to pioneers is given by equation (3) in the main text. All explanatory variables are measured in 2000 and are computed using 500 meters rings around each block (except the distance to subway, to parks, and to closest gentrifying block variables, as well as the waterfront dummy). Robust standard errors, corrected for cross-sectional spatial dependence within a 500 meters radius (using HAC estimation), are reported in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%. The proxy for "Socio-economic changes in the neighborhood 1990–2000" is the distance to the closest block that gentrified between 1990 and 2000 when our gentrification indicator is used as a dependant variable, and the change in average per capita income or in the share of educated residents between 1990 and 2000 for our two other proxies for socio-economic changes.

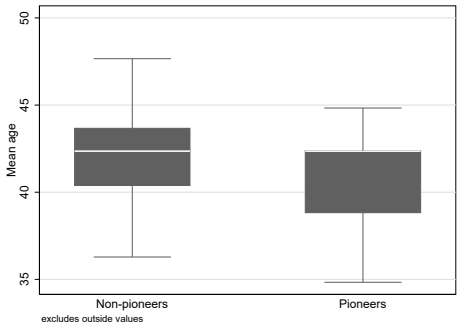
Table O.10: Interactions with prime locations and waterfront (500m), 2000–2010.

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
	(a) Interaction with prime locations					
Δ # pioneer estab.	1.168^{a}	0.195	1.835^{a}	0.883^{b}a	0.362^{b}	0.053
	(0.331)	(0.264)	(0.444)	(0.3332)	(0.149)	(0.106)
— × log dist. prime locations	0.217	0.516	- 0.185	0.342	0.027	0.083
	(0.281)	(0.327)	(0.383)	(0.350)	(0.119)	(0.128)
log dist. prime locations	-9.187	-108.340^{a}	-9.279	-100.618^{a}	-2.191	-36.510^{a}
	(6.055)	(17.366)	(6.950)	(19.374)	(2.573)	(6.680)
Controls	✓	✓	✓	✓	✓	✓
Observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.050	0.090	0.055	0.085	0.036	0.083
Sample	New York	NYC	New York	NYC	New York	NYC
	(b) Interaction with waterfront					
Δ # pioneer estab.	1.376^{a}	1.000^{a}	1.764^{a}	1.519^{a}	0.394^{a}	0.265^{a}
	(0.296)	(0.228)	(0.290)	(0.241)	(0.119)	(0.087)
— × log dist. waterfront	-0.309^{b}	-0.447^{b}	-0.569^{b}	-0.645^{b}	-0.078^{c}	-0.082
	(0.142)	(0.178)	(0.162)	(0.259)	(0.050)	(0.057)
log dist. waterfront	- 0.038	0.177	0.903	-0.295	-0.015	-0.032
	(0.796)	(1.192)	(0.757)	(1.047)	(0.237)	(0.315)
Controls	✓	✓	✓	✓	✓	✓
Observations	34,022	19,947	33,718	19,766	33,725	19,772
R-squared	0.051	0.078	0.056	0.079	0.036	0.073
Sample	New York	NYC	New York	NYC	New York	NYC

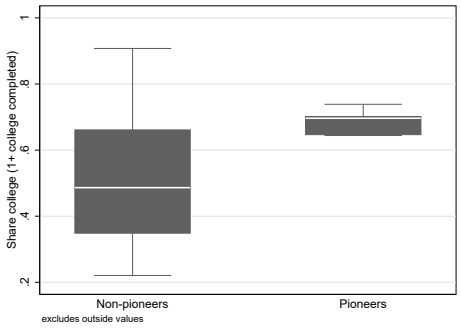
Notes: Reported coefficients and standard errors are mutiplied by 1,000 compared to the actual ones. All regressions include the following controls: Ln per cap. income; Share college edu. resid.; Ln rent; Median age of buildings; Share black resid.; Share asian resid.; Share other resid.; Ln population; Less than 200m from waterfront; Ln (1+# train lines); Ln (1+# bus lines); Ln distance to closest park; Ln # of main landmarks; Socio-economic changes in the neighborhood 1990–2000. For the sample limited to NYC, the controls also include # murder per cap.; # burglary per cap.; # robbery per cap.; # rape per cap.; Ln (1+# rent control buildings); Share vacant land; Presence of limited height districts; Presence of historical districts. Prime locations are provided by Ahlfeldt et al. 2020. The sample is composed of blocks with per capita income below the median in the city in 2000, and with at least eight residents. The measure of exposure to pioneers is given by equation (3) in the main text. All explanatory variables are measured in 2000 and are computed using 500 meter rings around each block (except the distance to subway, to parks, and to closest gentrifying block variables, as well as the waterfront dummy). Robust standard errors, corrected for cross-sectional spatial dependence within a 500 meter radius (using HAC estimation), are reported in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

Figure O.4: Selected characteristics of workers in pioneer sectors in the U.S., 2000–2010.

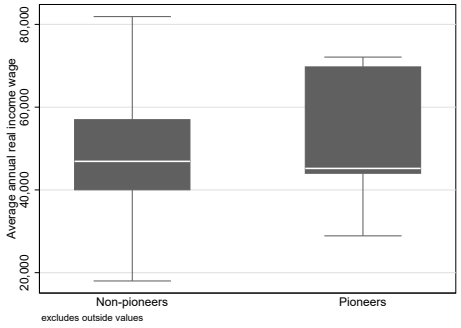
(a) are a bit younger.



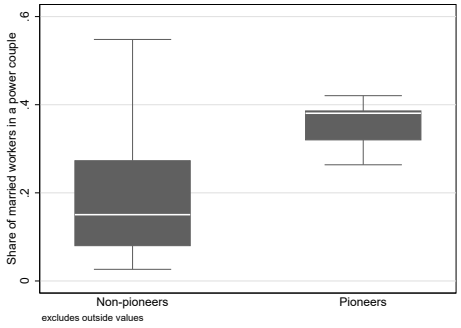
(b) are more educated.



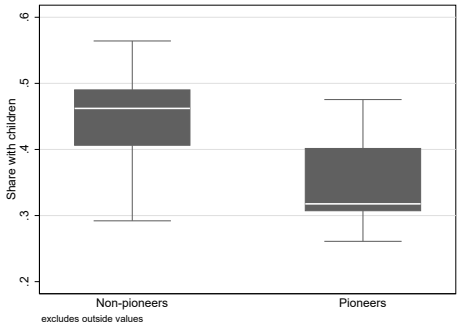
(c) are not paid much more.



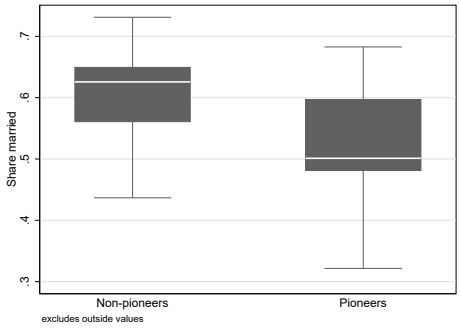
(d) are in 'power couples'.



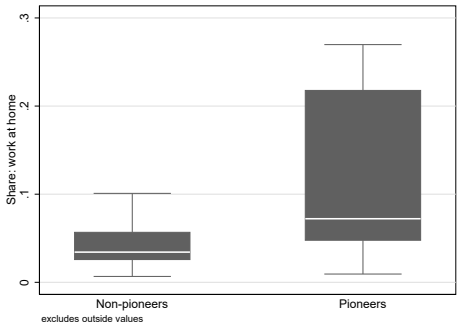
(e) have fewer children.



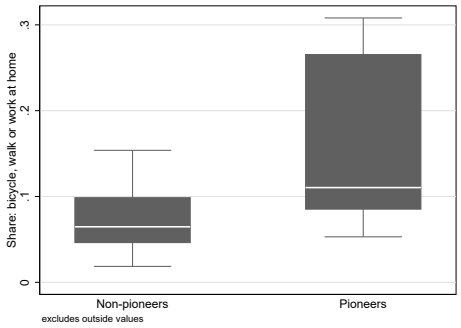
(f) are more often single.



(g) work more at home.



(h) commute shorter distances.



Notes: Our computations using IPUMS data for the U.S. for the years 2000 and 2010. Following Costa and Kahn 2000, we define power couples as couples in which both members are college educated.

Table O.11: Pioneers, creative sectors, and sectors with similar characteristics.

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
Δ # pioneer estab.	0.002^{a}	0.002^{a}	0.003^{a}	0.003^{a}	0.001^{a}	0.001^{a}
	(0.001)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)
Δ # other creative plants	-0.002^{c}	-0.002	-0.004^{a}	-0.003^{b}	-0.001^{a}	-0.001^{b}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)
Δ # plants with similar charact.	-0.004^{c}	-0.004^{c}	-0.004^{b}	-0.004^{c}	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Controls	✓	✓	✓	✓	✓	✓
Observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.052	0.078	0.057	0.080	0.037	0.075
Sample	New York	NYC	New York	NYC	New York	NYC

Notes: All regressions include the following controls: Ln per cap. income; Share college edu. resid.; Ln rent; Median age of buildings; Share black resid.; Share asian resid.; Share other resid.; Ln population; Less than 200m from waterfront; Ln (1+# train lines); Ln (1+# bus lines); Ln distance to closest park; Ln (1+# of main landmarks); Socio-economic changes in the neighborhood 1990–2000. For the sample limited to NYC, the controls also include # murder per cap.; # burglary per cap.; # robbery per cap.; # rape per cap.; Ln (1+# rent control buildings); Share vacant land; Presence of limited height districts; Presence of historical districts. The sample is composed of blocks with per capita income below the median in the city in 2000, and with at least eight residents. The measure of exposure to pioneers is given by equation (3) in the main text. All explanatory variables are measured in 2000 and are computed using 500 meter rings around each block (except the distance to subway, to parks, and to closest gentrifying block variables, as well as the waterfront dummy). Robust standard errors, corrected for cross-sectional spatial dependence within a 500 meters radius (using HAC estimation), are reported in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

O.10. Geography of gentrification

We here take a first look at the geography of gentrification in New York in 1990–2000 and 2000–2010. Starting with our discrete measure, we find that 3,259 (8.41%) of poor blocks (i.e. whose income per capita is below the median income per capita observed in the city) within a 30 kilometers radius around Wall Street are gentrifying during at least one of these two sub-periods. We identify 1,381 gentrifying blocks

between 1990 and 2000, and 1,878 between 2000 and 2010.[34] Quite naturally, only 20 blocks are identified as gentrifying in both periods as doing so entails very large socio-economic changes.

Figure O.5 depicts the geographic distribution of our three gentrification measures within the New York metro area, averaged over 1 kilometers rings centered on Wall Street. The left figure in panel (1) shows on the same graph the distribution of all blocks and of poor blocks. Compared to the distribution of all blocks, poor blocks are overrepresented in the 5–20 kilometer range in both periods, which suggests we should look beyond the most central parts of the city to study gentrification. The right panel in panel (1) shows that a large share of gentrifying blocks is concentrated in the 4–5 kilometer range where poor blocks are relatively abundant.[35] It further reveals that gentrification slightly shifted towards the more central parts of the city (Couture and Handbury, 2020). Observe there is little gentrification close to the center according to our discrete measure. The reason is that there are few poor blocks there. Hence, even if income growth remained strong in the center (see panel (2) of Figure O.5) it is hard to talk about gentrification in the usual sense.

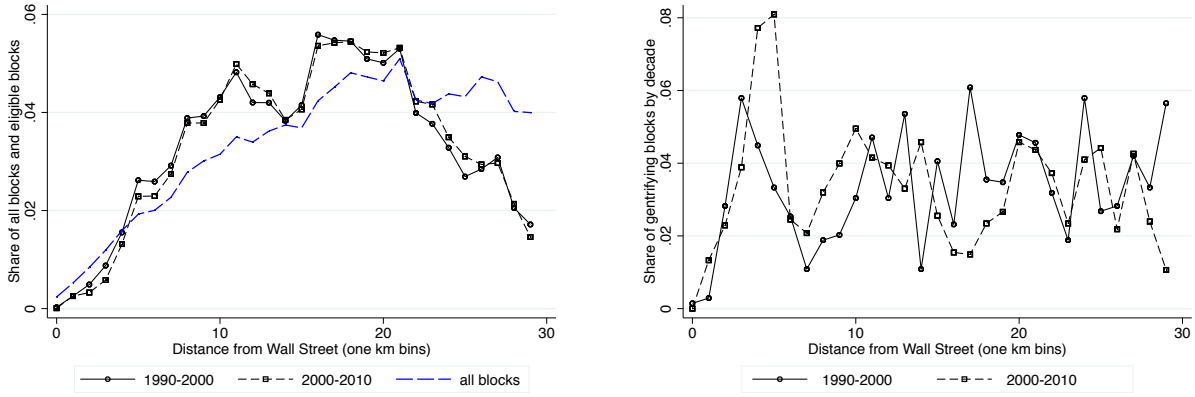
Panels (2) and (3) show that the geography of changes in income and education look markedly different. The left figures of panels (2) and (3) report the evolution of income and the share of educated for all blocks, whereas figures on the right focus on poor blocks. Whereas income changes are more skewed and concentrated in the most central parts of the city, education changes have been less skewed and

[34]The importance of gentrification has increased between 1990 and 2010. While 132,863 people lived in blocks that underwent gentrification in the former decade, 338,412 people lived in blocks that underwent gentrification in the latter decade.

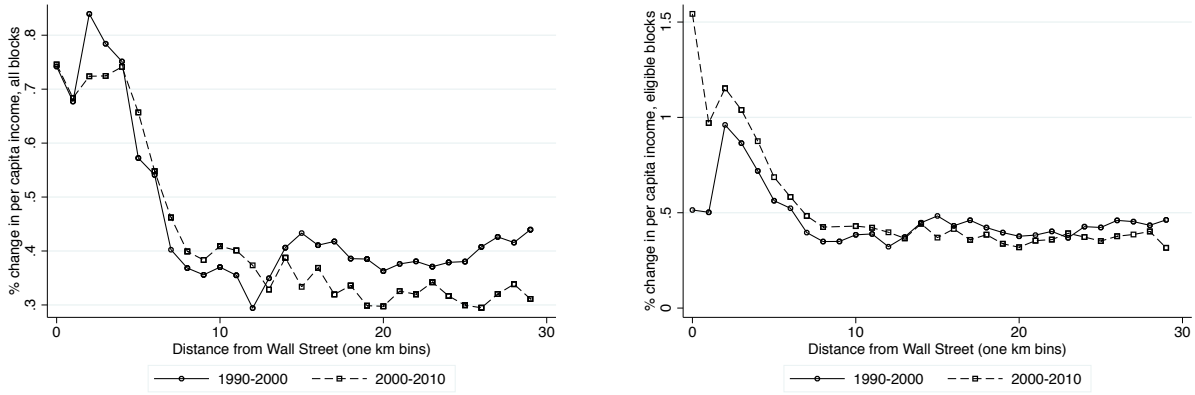
[35]Gentrification is not exclusively a central-city phenomenon: there are many gentrifying areas that are not close to the city center. A substantial amount of gentrification seems to partly follow the distribution of new housing, which is either located in the central city (due to renewal of old existing housing) or at the city fringe (due to construction of new housing; see Brueckner and Rosenthal 2009).

Figure O.5: Distribution of gentrifying blocks by decade and distance to Wall Street.

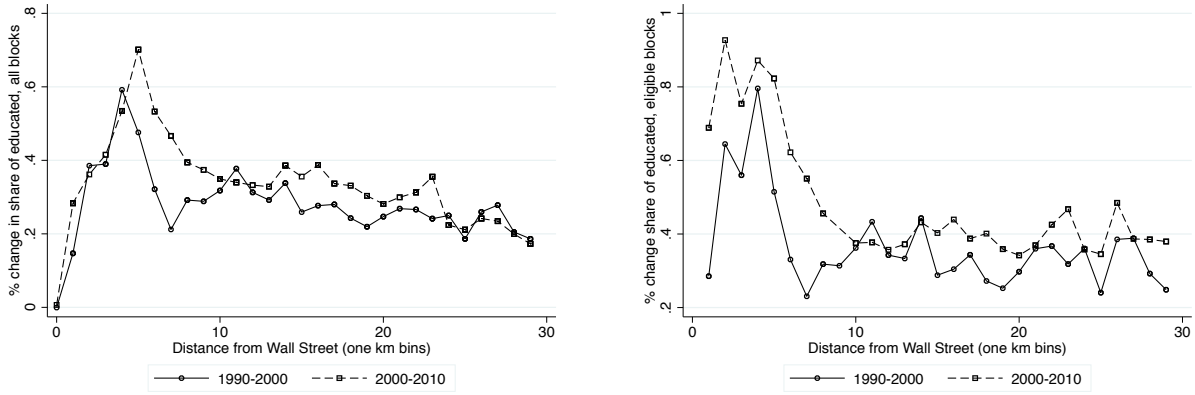
(1) Poor blocks and discrete measure of gentrification, γ_{it}.



(2) Continuous measure, percentage changes in ln per capita income.



(3) Continuous measure, percentage changes in the share of educated.



Notes: Distribution of gentrifying blocks within 30 kilometers around Wall Street. For the discrete measure of gentrification in panel (1), we report the sptaial distribution of blocks (all blocks and poor blocks) on the left, and share of blocks that gentrify (the total being all gentrifying blocks within 30 kilometers of Wall Street) on the right. For the continuous income and education measures, we report average changes by one-kilometer distance rings (panels 2 and 3). In panels (2) and (3), figures are for all blocks on the left and for poor blocks on the right. Income and education changes across blocks are trimmed by the bottom and top 1% of block-level distributions to remove outliers.

have moved slightly away from the most central parts. The block-level correlation between the percentage change in log income and the share of educated is only 0.24 in 1990–2000 and 0.11 in 2000–2010. It increases to 0.39 for poor blocks in 1990–2000, and jumps to 0.89 for poor blocks in 2000–2010. This shows that, before the 2000s, income and education changes were much less correlated, especially when it comes to gentrification.[36] Thus, looking beyond income as the sole criterion to understand gentrification, especially before 2000, seems warranted. The tighter link between income and human capital after 2000 echoes findings of the literature on the 'working rich' (e.g., Smith et al., 2019).

Additional references

Ahlfeldt, Gabriel M., Thilo N.H. Albers, and Kristian Behrens, "Prime locations," CEPR Discussion Paper DP15470, Centre for Economic Policy Research November 2020.

Bernard, Andrew B., Ilke van Beveren, and Hylke Vandenbussche, "'Concording EU trade and production data over time'," CEPR Discussion Papers 9254, C.E.P.R. Discussion Papers December 2012.

Carillo, Paul E. and Jonathan L. Rothbaum, "Counterfactual spatial distributions," *Journal of Regional Science*, 2016, *56* (5), 868–894.

[36] We confirm these results by regressing the percentage changes in log income on initial income, an indicator for poor blocks, and the percentage change in the share of educated (also interacted with the poor block dummy). While changes in the share of educated are positively and significantly related to income changes in both decades, it is particularly so in 2000–2010. In that period, the interaction with poor block is also positive, whereas it is not in the 1990s.

Costa, Dora and Matthew Kahn, "Power couples: Changes in the locational choice of the college educated, 1940-1990," *Quarterly Journal of Economics*, 2000, *115* (4), 1287–1315.

Martin, Julien and Isabelle Mejean, "Low-wage country competition and the quality content of high-wage country exports," *Journal of International Economics*, 2014, *93* (1), 140–152.

Neumark, David, Brandon Wall, and Junfu Zhang, "Do small businesses create more jobs? New evidence for the United States from the National Establishment Time Series," *The Review of Economics and Statistics*, August 2011, *93* (1), 16–29.

Pierce, Justin R. and Peter K. Schott, "Concording U.S. harmonized system categories over time," *Journal of Official Statistics*, 2012, *28* (1), 53–68.

Walls and Associates, "Technical documentation of the National Establishment Time Series database (NETS)," 2014.