Gentrification and pioneer businesses*

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

October 30, 2018

Abstract

We build a time-consistent block-level dataset from 1990 to 2010 containing socio-economic characteristics of residents and information on businesses in New York. We identify both gentrifying areas and pioneer sectors characterized by atypical location decisions. The latter—mostly cultural, recreational, and creative industries—help us to better predict and understand gentrification. We find that a block's initial exposure to pioneers has a quantitatively sizable effect in predicting future gentrification. Pioneers foster gentrification through the types of workers they hire, their signal as to the future prospects of a neighborhood, and their effect on the arrival of consumption amenities.

Keywords: Gentrification; pioneer businesses; microgeographic data; New York.
JEL Classification: R14, R23, R31.

*We 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 the 2016 NARSC Meetings in Minneapolis, the 2nd Kraks Fonds Workshop on Urban Economics in Copenhagen, the Workshop on 'Gentrification and Urban Neighborhood Dynamics' at UQAM Montréal, the 2017 Rotman-Sauder Conference in Real Estate and Urban Economics in Vancouver, the 2017 CEPR Conference on Urban and Regional Economics in Paris, the CMSSE winter workshop in Moscow, the Applied Micro Day in Montreal, the 2018 CEA conference in Montreal, the IEB 5th Urban Economics workshop in Barcelona, the 2018 UEA Conference in New York, and many seminars for constructive comments and suggestions. 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. This study was funded by the Russian Academic Excellence Project '5-100'. Any remaining errors are ours.

†ESG, Université du Québec à Montréal, Canada; National Research University Higher School of Economics, Russian Federation; and CEPR, UK. E-mail: behrens.kristian@uqam.ca

‡ESG, Université du Québec à Montréal, Canada. E-mail: boualam.brahim@uqam.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 Université catholique de Louvain, Belgium. E-mail: mayneris.florian@uqam.ca

1 Introduction

After four decades of suburbanization, central cities have started to regain population in the 1990s, becoming increasingly attractive to young, educated, and wealthy people. This process of *gentrification* has accelerated in the 2000s and profoundly altered the social and economic fabric of many inner cities. While much work has focused on residents' socio-economic characteristics in central cities broadly defined, changes at smaller geographic scales and their interactions with local businesses have attracted less attention. Our aim is to provide systematic quantitative evidence on the very localized nature of gentrification and the role that local businesses play in that process. To this end, we use detailed microgeographic data to detect gentrification at the block level and to identify *pioneer* sectors that are overrepresented in areas that will gentrify in the future. These pioneers—which are concentrated in the cultural, recreational, and creative industries—have a quantitatively sizable effect in predicting future spots of gentrification and seem to play an important role in the subsequent gentrification process. We find that they may affect gentrification through the type of workers—young, educated, without children, and in power couples—they hire, the signal they produce regarding the future prospects of the neighborhood, and their effect on the subsequent arrival of consumption amenities valued by young and educated residents. Our results thus show that at the early stage of neighborhood change, the presence of artists, architects, art dealers and other creative businesses heralds future gentrification. Consumption amenities arrive at a later stage of the process. Hence, the usual suspects—such as Starbucks, Whole Foods, and Trader Joe's—follow rather than lead the pioneer businesses we identify

Geographers, sociologists, and urban planners have extensively investigated the dynamics of specific neighborhoods.[1] More recently, economists have examined gentrification from a more quantitative perspective. The vast majority of these studies try to understand *why* gentrification has spread across American inner cities over the last decades.[2] While this question is surely important, analyzing *where* gentrification occurs and *how* it emerges is also key to understand this phenomenon, both from academic and policy perspectives. Hwang & Lin (2016) point out that gentrification arises at a very localized scale, perhaps the block- or even the building-level. This is also one key insight by Easterly et al. (2018) who retrace the changes that took place in one New York City block (Greene Street) over 400 years. They emphasize that "surprises"—localized unanticipated shocks—are key to understand the block's fate. These "surprises" often include the decisions of businesses to start operating in the area—in their case study, e.g., the garment industry (1880–1890), and artists and art galleries (1960–1970).

Given the importance of idiosyncratic shocks and the need for a microgeographic scale,

[1] See, e.g., Lees (2003) and Zukin et al. (2009).

[2] See Guerrieri et al. (2013), Edlund et al. (2015), Hwang & Lin (2016), Baum-Snow & Hartley (2016), Couture & Handbury (2017), and Couture et al. (2018).

analyzing in a systematic way *where* gentrification occurs and *how* certain types of businesses contribute to it is challenging. To make progress in this direction, we tackle three key challenges in our analysis. Our first challenge is to build a dataset containing socio-economic characteristics of residents and information on businesses at a very fine-grained geographic scale. To this end, we combine information on residents and housing at the block- and block-group level in the New York metropolitan area with geocoded data on the establishments operating in this area between 1990 and 2010. We develop a concordance algorithm to obtain a stable and highly disaggregated geography that spans the two decades and that allows us to detect gentrification at small spatial scales. Our final dataset contains decennial block-level information for the New York metropolitan area between 1990 and 2010, including their demographic characteristics, their housing characteristics, the quality of their local environment (crime, amenities), and their composition in terms of businesses. We use that dataset to identify blocks that gentrify between 1990–2000 and between 2000–2010.

Our second challenge is to identify the types of businesses that precede gentrification. All current studies that analyze the links between businesses and neighborhood change focus on *ad hoc* lists of business types they consider being important.[3] We adopt a different strategy and exploit *atypical location patterns* of establishments. To this end, we estimate count models and show that some sectors that are usually found in affluent areas are also overrepresented in poor blocks that will gentrify in the subsequent decade. We refer to these sectors as *pioneers*. We identify 21 pioneer industries out of 430 6-digit NAICS industries in New York over the decade 1990–2000. Our list includes, among others, architects, designers, or art dealers', the latter having often been associated with gentrification episodes in the literature (e.g., Schuetz 2014; Easterly et al. 2018). Although we use data from New York in 1990–2000 to detect pioneers, we show that they are associated with gentrification more broadly, both in subsequent decades (New York between 2000 and 2010) and in other cities (Philadelphia). Controlling for a large set of initial characteristics of the blocks, the block's *exposure to pioneers* is positively related to the probability that it will gentrify during the next decade. This effect survives a battery of tests and is robust to IV estimations and corrections for spatial correlation. It is also quantitatively sizable: it increases the share of blocks that are predicted as gentrifying and that are within 100 meters distance from blocks effectively gentrifying by more than 20% in New York, and by about 18% in Philadelphia. In a nutshell, the location of pioneers carries information as to subsequent gentrification.

Our third challenge is to better understand the possible mechanisms underlying the association between pioneers and gentrification. We discuss several explanations that are not mutually exclusive. First, we show that workers employed in pioneer industries tend to have characteristics usually associated with "gentrifiers": they are young, educated, single or matched with

[3]Schuetz (2014) looks at art galleries, while Couture & Handbury (2017) define a list of sectors that are consumption amenities for the young and educated—including restaurants, bars, gyms, and personal services.

a highly educated partner, and they tend to live close to their workplace. This points to a tight connection between the mix of businesses in a given area and its subsequent evolution in terms of residents. Second, we find indirect evidence of a possible signaling effect. Since pioneers are usually found in more expensive areas, their presence in poor blocks—an atypical location pattern—may signal future neighborhood improvements (see, e.g., Caplin & Leahy 1998). Finally, we find no evidence that pioneer industries are consumption amenities *per se* for local residents. However, their initial presence is associated with the subsequent arrival of various consumption amenities such as restaurants, bars, gyms, and personal services that are highly valued by young and educated workers (see Couture & Handbury 2017).

Our paper contributes to an abundant literature on neighborhood change, including central-city revival, neighborhood tipping, urban renewal, urban decay, and employment decentralization.[4] It also contributes to the economic literature on gentrification, a specific type of neighborhood change.[5] Our contribution is twofold. First, we put establishments at the core of our quantitative analysis. Previous works on gentrification focus mostly on demographic changes, amenities, and housing market dynamics in gentrifying areas, yet are basically silent on the composition of, and changes in, stores, businesses, and economic activity in these neighborhoods.[6] Only a few papers link gentrification with businesses in a quantitative framework. Lester & 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 & Handbury (2017) investigate the role of consumption amenities in driving gentrification in central cities. Glaeser et al. (2018) look at the change in the number of grocery stores, cafés, restaurants, and bars during the gentrification process using Yelp data at the ZIP code level. They show that coffee shops help predict gentrification.[7] Contrary to previous work, we

[4]See, e.g., Glaeser & Kahn (2001), Burchfield et al. (2006), Duranton (2007), Rosenthal (2008), and Lee & Lin (2018). See Rosenthal & Ross (2015) for a recent state-of-the-art survey.

[5]O'Sullivan (2005) and Ellen et al. (2017) discuss the possible circular relationship between gentrification and crime. Brueckner & Rosenthal (2009) highlight the age of the housing stock as a driver of gentrification. McKinnish et al. (2010) and Ding et al. (2016) show that gentrification is, on average, not associated with the displacement of poor or minority incumbent residents. Edlund et al. (2015) highlight the role played by full-time educated workers—whose tolerance for commuting decreased over the past decades—in driving up housing demand in central neighborhoods. Su (2018) points at the role of the rising value of time of highly-skilled workers. Baum-Snow & Hartley (2016) and Couture & Handbury (2017) document the urban revival process at work in U.S. cities since the 1990s and discuss whether this revival is mainly due to changing amenities or changing valuation of these amenities by specific groups of residents. Finally, Couture et al. (2018) and Brummet & Reed (2018) investigate the welfare implications of gentrification.

[6]Some interesting exceptions exist among geographers and sociologists (see, e.g., Lees 2003, Zukin et al. 2009, or Sullivan & Shaw 2011 on retail gentrification), but this strand of literature largely relies on qualitative case studies, thereby loosing the strength of quantitative analysis.

[7]Glaeser et al. (2018) measure gentrification using changes in housing prices. In our data, the surge in prices follows the change in the demography of a neighborhood. We thus focus on an earlier stage of gentrification as

propose an empirical methodology to estimate which sectors are associated with subsequent gentrification at its very early stages, when the blocks are not yet gentrified from a residential point of view.

Our second contribution pertains to the geographic level of the analysis. The extant literature on intra-city dynamics relies mostly on data for zip codes or census tracts (e.g., Glaeser & Kahn 2001, Echenique & Fryer 2007, Rosenthal 2008, McKinnish et al. 2010, Baum-Snow 2014). This spatial scale is probably sufficiently detailed to document phenomena like urban renewal or the revival of city centers. Yet, in line with the observations by Hwang & Lin (2016) and Easterly et al. (2018), this scale may be too large to capture gentrifiction. We show that gentrification dynamics indeed take place at very small geographic scales that hardly follow census tract or zip code boundaries. This is also consistent with recent work on the determinants of social mobility by Chetty et al. (2018) who point out at that the "neighborhoods" that matter for child outcomes are "hyperlocal" and clearly smaller than census tracts.[8] In our case, we show that working with census tracts or zip codes changes the areas that are detected as gentrifying, and also modifies the list of pioneers we estimate. The use of fine-grained data allows for a more precise delineation of the gentrification process, thereby reducing measurement error. Using such fine-grained data we can show, for example, that gentrification is not as central as one would expect: in New York, gentrifying areas in 1990–2000 are located, on average, 12 kilometers from Wall Street.

The remainder of the paper is organized as follows. We present our data and describe the methodology used to produce stable geographic units and to identify gentrifying blocks in Section 2. We document the geography of gentrification in New York in Section 3. We identify pioneers and assess their predictive power to detect future spots of gentrification in Section 4. In Section 5, we investigate why pioneers may herald gentrification. Last, Section 6 concludes. We relegate parts of the data description, technical details, and additional tables and figures to a set of appendices.

2 Data

We first provide a detailed description of the data we use in our empirical analysis of gentrification in New York between 1990 and 2010. Additional information, including on comparable data that we use for Philadelphia to complement our analysis, is relegated to Appendix A.

compared to them, which may explain why coffee shops do not show up in our list of pioneers.

[8]For example, Chetty et al. (2018) show that the poverty rate of blocks surrounding the place where the child grew up is a strong determinant of the child's future outcomes. This holds, however, only true for blocks within about half a mile, suggesting that neighborhood effects are highly localized.

2.1 Census data

Contrary to previous studies on segregation or neighborhood change that work with time-consistent census tract data—usually the Geolytics Neighborhood Change Database— we work at a finer geographic scale using time-consistent 'blocks'. To this end, we first extract census block-level data from the National Historical Geographic Information System (NHGIS) of the population center at the University of Minnesota. We gather information on population and housing in 1990, 2000, and 2010. These include the number of residents, their demographic characteristics, and information on the age of the housing stock. Census blocks are the most spatially disaggregated census unit, with more than 200,000 blocks in the New York metropolitan area. Counts of residents and housing units are directly available at the block level. Several other variables—such as total income or the number of residents by educational attainment or by race—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.

Going from block groups to blocks has two main advantages. First, the boundaries of blocks and block groups change between census years; but contrary to blocks, there are no relationship files linking block groups over time. Hence, as shown below, the time-consistent geography for census blocks is much finer than that obtained for block groups. Second, and as a consequence of the former, block-level information allows us to work at a much finer spatial scale than tracts or block groups do. We discuss in Section 3.3 the importance of working at the finest possible geographic scale when dealing with gentrification.

Concording census blocks. The number and boundaries of census blocks—and of other census geographic units—change over time. In the NY metropolitan area, the number of census blocks increased from 189,976 in 1990 to 240,318 in 2010.[9] This increase masks a wide range of changes made by the Census to the geography of these blocks. Some are split into several new blocks, while others are grouped together. More problematically, some blocks are split and their parts are recombined in complex ways into several new or existing blocks. To deal with these problems, we have developed a concordance methodology that can be used to create constant geographies based on census blocks (see Appendix B for details).[10]

[9] See Appendix A.1 for a more precise definition of what we refer to as NY metropolitan area.

[10] The novelty of our method is to view the concordance problem from a graph-theoretic perspective. The algorithm builds on the observation that determining the optimal concordance (i.e., the smallest synthetic groups) simply corresponds to finding the connected components of the graph spanned by the initial blocks and the revisions. 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

Our methodology creates 'smallest synthetic groups'. As is well known, these 'smallest synthetic groups' are by definition more aggregated than the original units. This naturally raises the question of how much geographic information we lose when concording census blocks over time. Our algorithm identifies 152,529 time-consistent *concorded blocks* (henceforth blocks, for short) for New York over 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.

Table 1: Number of census blocks per concorded blocks, 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
# of census blocks, 30 km around Wall Street	1	1	1.7	2	4.3	74.7	1

Notes: This table reports the distribution of the average number of census blocks per concorded block in New York for 1990, 2000, and 2010. There is a total of 152,529 concorded blocks in our dataset. A number of census blocks equal to 1 in the table means that, on average, over 1990–2010 the concorded block consists of just 1 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 'central part' of the city, defined as being within a radius of 30 kilometers around Wall Street.

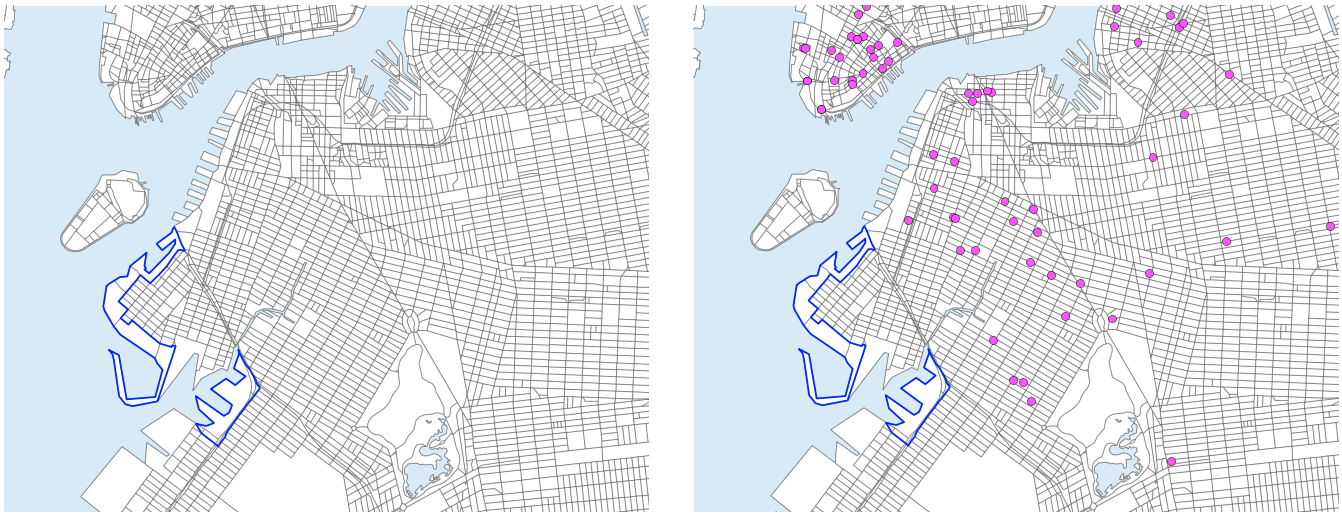
Table 1 summarizes the distribution of the number of census blocks that are contained in our concorded blocks. As shown, more than 75% of our blocks consist of a single census block, more than 90% are made of 1 or 2 census blocks, and more than 95% are made of 1 to 3 census blocks only. Less than 1% of our concorded blocks contain more than 8 census blocks. On average, a 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).

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 (see Table 1). 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. As can be seen from panel (a) of Figure 1—which illustrates the relationship between our concorded blocks (blue polygons) and original census blocks (grey polygons) in the area known as 'Red Hook'—some areas along the waterfront or parks change substantially between census years. Since anecdotal evidence suggests that many developments along the

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 & Rothbaum 2016 for an application to Washington DC), whereas our method deals with all years simultaneously.

Figure 1: Concorded blocks versus census blocks and the location of art dealers.

(a) Illustration of concorded blocks. (b) Location of art dealers.



Notes: Concorded blocks, 1990–2010, as determined by the methodology explained in Appendix B.

waterfront correspond to gentrification, it is especially important to have a constant geography there to precisely capture demographic and socio-economic changes over time.

One final comment is in order. Since some of our census variables are reported at the block-group level only, one may ask why we do not directly work at that level. As stated above, there are no census relationship files for block groups, which makes any concordance problematic. 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 concorded 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 concorded block group gathers all block groups bordering both the Hudson river and the East river in Manhattan. Similar problems arise for Staten Island. Such a coarse stable geography makes the identification of highly localized dynamics extremely difficult.

2.2 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 (henceforth, NETS). We use the 2014 version of that dataset for the New York core-based statistical area (CBSA) 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, total employment at the establishment, and several credit score indices.

Table 2: Sectoral breakdown of the NETS data for the New York metro area.

	(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. Panel (b) reports the number of establishments by establishment size category, as well as the total number of establishments. All computations based on the NETS New York CBSA dataset.

One important feature of the NETS data for our purpose is the location information of the establishments. Depending on the precision of the geocoding, the latitude and the longitude reported in the NETS data are mainly based on either block face ('rooftop') or zip-code. "Block face" 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 is more precise than, e.g., census tracts). Panel (a) of Table 2 summarizes the accuracy of the geocoding in our dataset. It shows that three-quarter of the establishments, accounting for 77% of employment, are geocoded at block face 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 2 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 million in 2010. This increase is entirely driven by the surge in the number of self-employed and the increase in the number of establishments with 2 to 5 five employees. We discuss this issue further in Appendix A and explain why this is not a problem for our analysis.

Assigning establishments to blocks. We use geographical information system (GIS) software to assign the NETS establishments to our 152,529 concorded blocks for the years 1990, 2000, and 2010, 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 census blocks of the New York metro area. Panel (b) of Figure 1 illustrates the granularity of our data. It depicts the location of art dealers (NAICS 453920) in northern Brooklyn and southern Manhattan in 2000.

2.3 Blocks with demographic characteristics and establishments

Our final dataset collects information on the demographic and socio-economic composition (population, race, income, educational attainment) of the concorded blocks, constructed from the census data, and information on the economic activities located in the blocks (number of establishments, jobs, and sales by 6-digit industries) constructed from the NETS data. We further have information on the housing stock, its age distribution, as well as housing values and rents.

Table 3: Characteristics of concorded blocks, 1990–2010.

	1990				2000				2010			
		Percentile				Percentile				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 concorded blocks whose centroids are less than 30 kilometers from Wall Street and which are not exclusively composed of water. There are 63,374 such concorded blocks in total. All monetary values are expressed in current U.S. dollars, except median housing values which are expressed in thousands.

Since gentrification is generally perceived as being a central city phenomenon, we focus in the remainder of the paper on blocks whose centroids are located less than 30 kilometers from Wall Street (which we take as the city center, following Glaeser & Kahn 2001). This leaves us with 63,799 blocks—62,168 of which have at least 8 inhabitants in 1990. These blocks represent around 60% of the population, establishments, and jobs in the New York MA over our study period.[11]

Table 3 summarizes basic characteristics of our blocks. They have on average roughly 187 residents over the period 1990–2010. The population distribution across blocks is quite skewed since the median number of residents is only about half of the average number. This skewness is—unsurprisingly—even more striking for the number of establishments and jobs, in line with the well-documented fact that economic activity generally displays even more geographic concentration than population. This contrasts with the distributions of per capita income and the share of educated people—defined as those with at least some college degree—where the median is not far from the average. Finally, we observe a dramatic increase in housing prices over the period. In twenty years, median housing values increased by about 150%,

[11]Other recent studies analyzing urban renewal and central city revival use different and arguably more restrictive definitions of the central city. Baum-Snow & Hartley (2016) take a 4 kilometers radius around Wall Street in their analysis, which captures about 2.8% of the metro population in our case. Couture & Handbury (2017) choose a variable distance cutoff to capture 5% of the metro population. In our case, this would correspond to a 5.5 kilometers radius around Wall Street.

while median gross rents almost doubled. This change is especially pronounced between 2000 and 2010.

2.4 Other datasets

We supplement the NETS and census data with several other datasets. These datasets are used to create additional controls for our regressions. More information is relegated to Appendix A.

Crime. We obtain information on crime from the Furman Center for Real Estate and Urban Policy. Our crime data is reported at the precinct level and is, therefore, gathered only for the five boroughs of New York City.[12] 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.

Public transportation. The location of subway stations, provided by the Metropolitan Transportation Authority (MTA), is obtained from the NYC OpenData website. For 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.

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.

Geographic controls. We complement our dataset with various 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 simple 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

[12] Few missing values are filled with similar indicators from the *Historical New York City Crime Data* provided by the New York City Police Department.

lacking a description). We further use GIS software to create a 'waterfront' variable, which is the distance of each block to the closest block that is composed exclusively of water.

3 Gentrification: Definition and geography

The aim of this section is to substantiate a number of basic facts related to the geography of gentrification. After having presented and discussed our definition of this phenomenon, we show that it is very localized and that data at the finest spatial scale are required to adequately measure it.

3.1 Identifying gentrifying blocks

It is fair to say that there is no consensus in either economics or sociology on how to quantitatively define gentrification. As emphasized by Barton (2016, p.3), this *"lack of consensus concerning the conceptualization of gentrification allowed researchers to identify gentrified neighborhoods in a variety of ways."* Consequently, the literature abounds with different quantitative definitions of gentrification. A series of papers use the (absolute or relative) change in income in a neighborhood as the single criterion of gentrification.[13] Another series of papers use multivariate strategies—related mostly to income, educational attainment, or housing prices—to define gentrification (see Barton 2016, for a discussion and comparison of the methods).[14]

The leitmotif underlying all these papers is that gentrification is a process affecting initially rather poor neighborhoods which experience a substantial influx of wealthier and more educated residents over a given period of time (e.g., Freeman & Braconi 2004, Zukin et al. 2009, McKinnish et al. 2010, Lester & Hartley 2014). Viewed in that way, gentrification is *"a dramatic shift in [the neighborhood's] demographic composition toward better educated and more affluent residents"* (Freeman & Braconi 2004, p.39). In line with that definition, we use three characteristics to identify blocks that gentrify over a ten year period: (i) a low initial level of per capita income; (ii) a substantial change in per capita income; and (iii) an accompanying change

[13]For example, neighborhoods whose median income is less than 50% of the MSA median income, and that move above this threshold in the next period, are defined as gentrified by Hammel & Wyly (1996). Similarly, gentrifying neighborhoods are defined by Meltzer (2016) and Meltzer & Ghorbani (2017) as census tracts that start in the bottom quintiles of the income distribution and that experience an increase in their income relative to the rest of the MSA. McKinnish et al. (2010) define gentrification as low-income neighborhoods (i.e., neighborhoods in the first quintile of the average family income distribution) which experienced a rise in this average income by at least 10,000$ between 1990 and 2000.

[14]Freeman (2005) defines gentrifying areas as initially disadvantaged neighborhoods which experience a rise in educational attainment larger than the median of the CBSA, and a rise in real housing prices. Similarly, Hammel & Wyly (1996) and Bostic & Martin (2003) use a scoring technique based on nine criteria to identify gentrifying neighborhoods, including the change in educational attainment, family income, type of jobs, and the proportion of residents aged 30 to 34.

in the share of highly educated residents. More precisely, we first rank the blocks based on their per capita income.[15] We then impose four conditions for a block to be considered as gentrifying: (i) there must be a minimum number of residents in the block (eight in our case) for the measured variations to be meaningful; (ii) the block must be relatively poor initially, i.e., below the metropolitan median per capita income; (iii) the block must move at least three deciles upwards in the metropolitan per capita income distribution over a ten year period; and (iv) the block must move at least one decile upwards in the metropolitan distribution of the share of educated residents over the same period. We use no other data—especially on racial composition, housing prices, establishments, and jobs—to identify gentrifying areas.

We run the exercise for two periods: 1990–2000 and 2000–2010. Given our definition of gentrification, 8.35% of blocks within a 30 kilometers radius around Wall Street—and below the median in terms of per capita income in the metropolitan area in 1990 or 2000—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. Quite naturally, only 20 blocks are identified as gentrifying in both periods as doing so entails huge socio-economic changes.

3.2 Where did gentrification occur in New York?

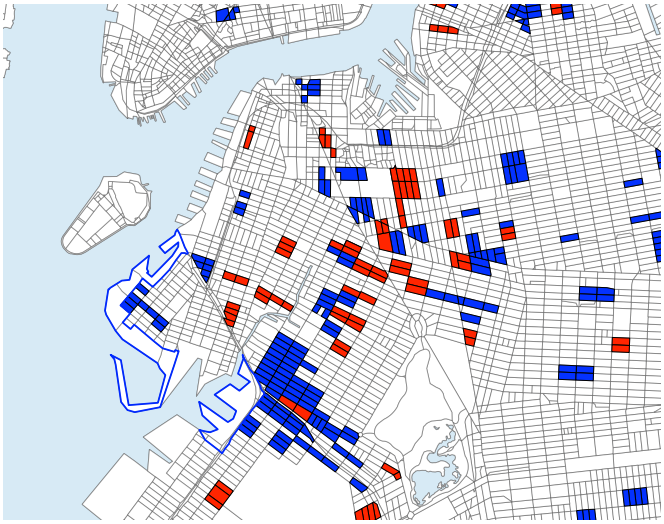
Given our definition of gentrification, where did it take place in New York between 1990 and 2010? Figure 2 depicts the gentrifying blocks we identify in northern Brooklyn in 1990–2000 (in red) and in 2000–2010 (in blue). It reveals a somewhat sequential pattern of gentrification: blocks that gentrified between 2000 and 2010 were likely to be 'close' to blocks that gentrified between 1990 and 2000. Observe that while there generally is 'contagion'—a spatial diffusion process—in gentrification, there is also much new gentrification in areas that are remote from previously gentrifying locations. As argued in the introduction, idiosyncratic changes—"surprises" in the terms of Easterly et al. (2018)—also seem to be important in determining the fate of individual areas.

To draw a map of gentrification—and to more closely link that map to the many stories reported in press articles—it is useful to delineate gentrifying 'neighborhoods' using the gentrifying blocks we identified. To do so, we run a cluster procedure to detect hot spots of gentrifying blocks (see Appendix C for details).[16] We identify 118 gentrifying neighborhoods

[15] This ranking is based on the entire metropolitan area and not only on the blocks within a 30 kilometers radius around Wall Street. Indeed, due to spatial sorting the households living closer to the central city are certainly not representative of the households in the entire metropolitan area.

[16] Our neighborhoods are built by detecting focal gentrifying blocks and putting buffers around those focal blocks. We ignore in this section a small number of isolated blocks (30 out of 1,378 gentrifying blocks between 1990 and 2000, and 44 out of 1,878 between 2000 and 2010) where rapid socio economic change occured but where we cannot really talk about neighborhoods in a meaningful sense. These blocks will, however, be kept for the rest of the empirical analysis.

Figure 2: Example of gentrifying blocks in northern Brooklyn.



Notes: Gentrifying blocks as identified using our criteria. Red blocks gentrified between 1990–2000 and blue blocks between 2000–2010.

for 1990–2000, and 135 gentrifying neighborhoods for 2000–2010 in a 30 kilometers radius around Wall Street. The maps of gentrifying neighborhoods for the two periods are provided in Figure 3.

Table 4: Top-10 gentrifying neighborhoods by population.

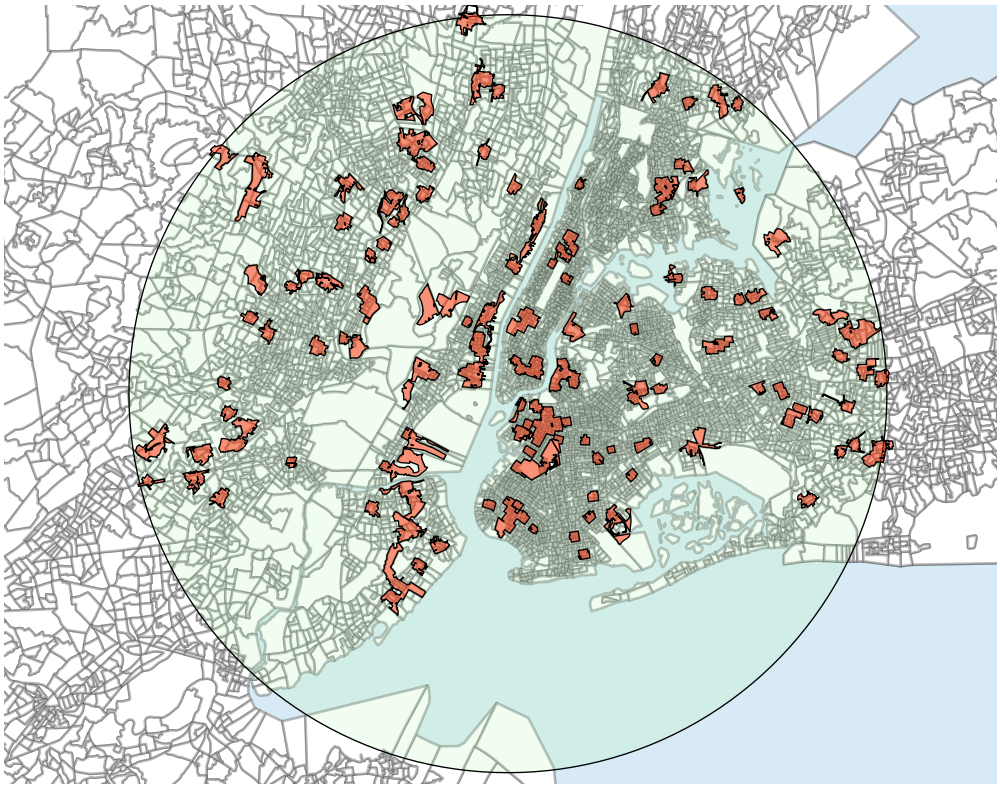
Gentrification 1990–2000		**Gentrification 2000–2010**	
Neighborhood	Population, 1990	Neighborhood	Population, 2000
Lower East Side	105,879	Harlem-Riverside Drive-Columbia U	234,699
Fort Greene-Lafayette-Bergen	77,316	Fort Greene-Union Street-Prospect Avenue	164,069
Morris Park-Westchester	54,281	East Village-Bleeker Street-Columbus Park	133,754
Hells Kitchen-Chelsea-Garment District	41,112	Astoria	106,533
Williamsburg	39,425	Williamsburg	88,406
Dyker Heights	37,696	Bay Parkway	87,541
Central Harlem (135th street)	27,487	Flushing Meadows-Rego Park	86,842
South-central Harlem (118th street)	26,549	Inwood Hill Park-Washington Heights	75,124
Upper East Side (96th Street)	23,147	Bedford-Stuyvesant-Bushwick	60,000
Hoboken-Jersey Shore	22,587	Downtown Jersey City-Paulus Hook	37,390
Total number of gentrifying neighborhoods	118		135
CBSA total population (<30km around Wall Street)	10,651,102		11,535,382
Pop. in gentrifying neighborhoods (share CBSA pop.)	805,814 (7.57%)		1,444,253 (12.52%)
Pop. in gentrifying blocks (share CBSA pop.)	132,863 (1.25%)		338,412 (2.93%)

Notes: Neighborboods as identified by our statistical procedure outlined in Appendix C. These neighborhoods do not coincide with administrative jurisdictions, and their naming is thus somewhat arbitrary. When a neighborhood is well identified, we use its common name (e.g., using the New York Times designation). If not, we choose a name representative of the area. All population figures used are for the initial year (i.e., 1990 for 1990–2000, and 2000 for 2000–2010).

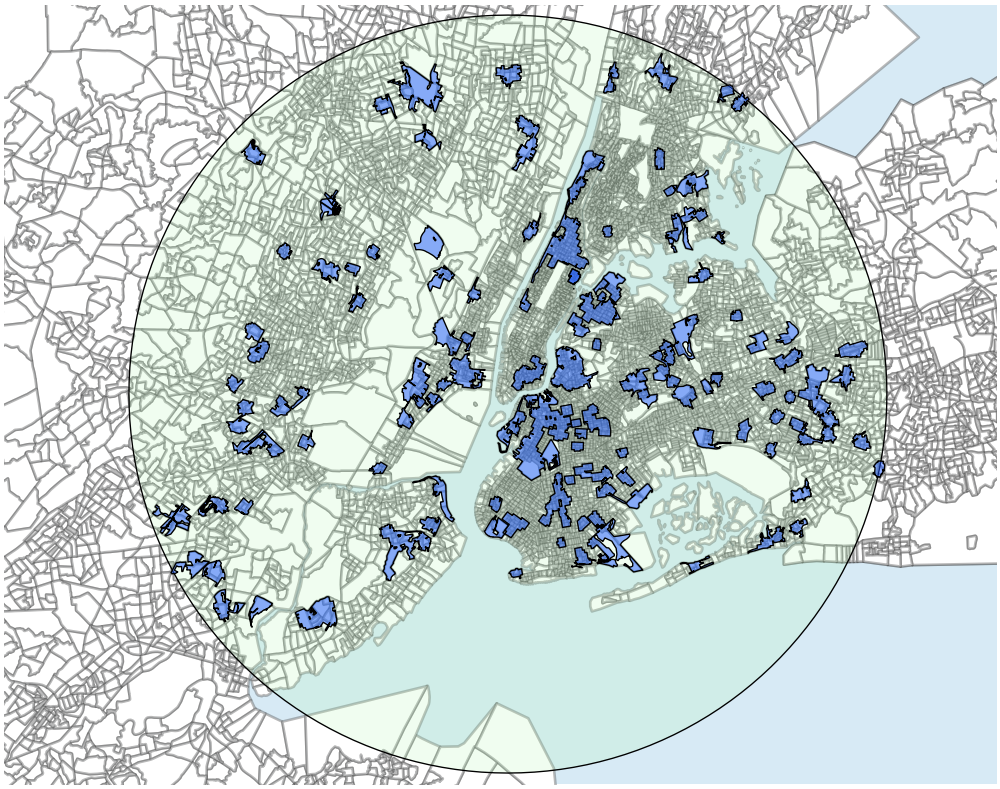
Table 4 reports the 10 most populated gentrifying neighborhoods in each of our two decades. Most of them are famous examples of gentrification: Fort Greene, Williamsburg, East Village, Lower East Side, Chelsea, and Harlem feature everywhere—from the New York Times' articles on gentrification to the popular travel guides. More interesting are the neighborhoods that most non New Yorkers are not familiar with. Some of them make our top 10. Astoria, for in-

Figure 3: Gentrifying neighborhoods in New York at less than 30 kilometers from Wall Street.

(a) 1990–2000.

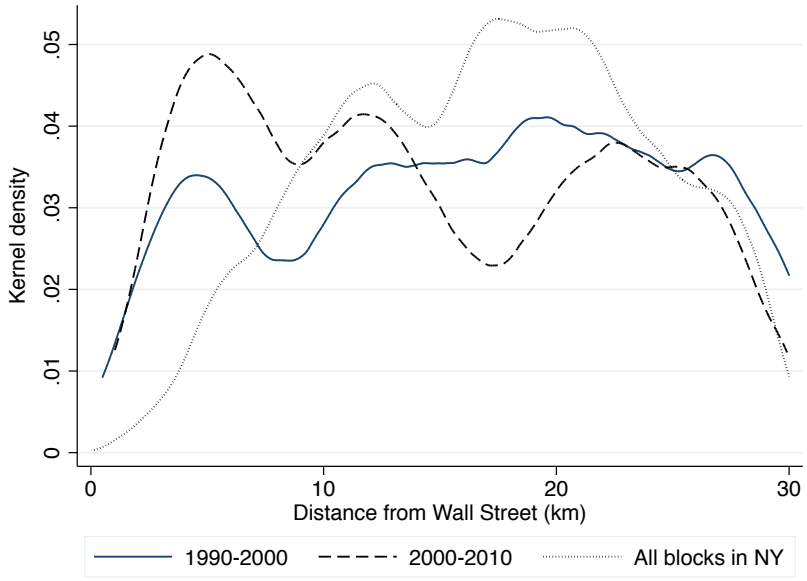


(b) 2000–2010.



stance, has gentrified quite abruptly between 2000 and 2010. It is presented as a new hot spot for young professionals and the most gentrified neighborhood in Queens.[17] Dyker Heights is a neighborhood located in South Brooklyn. This area is one of the safest neighborhoods in the city according to CompStat and is known for the specific design of its residences. Other gentrifying neighborhoods that are less populous and well known include Inwood (north of Manhattan along the Hudson river) or part of the Bronx (south of Morris Park, also known as the "new Little Italy") for example. However, gentrification is not limited to the five boroughs of New York. Locations on the Hudson waterfront like Hoboken, or the Village in Jersey City have also experienced episodes of gentrification according to our measure. This is again consistent with ample anecdotal evidence.

Figure 4: Distribution of gentrifying blocks by distance from Wall Street.



Notes: Distribution of all blocks and of gentrifying blocks around Wall Street.

Table 4 and Figures 3 and 4 highlight three facts. First, the magnitude of the gentrification phenomenon has increased between 1990 and 2010. While 'only' 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. In other words, the scale of gentrification has increased. Second, gentrification has gradually shifted more strongly towards central locations. As Figure 4 shows, gentrifying blocks are closer to Wall Street in 2000–2010 than they were in 1990–2000. Yet, although gentrification takes place mostly in more central parts of the city, it is not exclusively a central city phenomenon: there are many gentrifying areas that are not that close to the city center.[18] Disregarding those areas in the analysis of gentrification—by

[17] See, e.g., `https://macaulay.cuny.edu/seminars/rosenberg09/articles/t/h/e/The_Gentrification_of_Astoria_3a1c.html` for a detailed account.

[18] One example is Jamaica, which is close to JFK airport and identified as a gentrifying area with our definition.

focussing solely on a narrowly defined central city—may be unwarranted.

3.3 Does the spatial scale matter?

Most previous research on urban neighborhood change has relied on tract-level data.[19] One may thus naturally ask whether our block-level data capture the same information as tract-level data do. Figure 5 depicts the 2010 census tracts (gray lines) and our gentrifying blocks (red and blue zones) for two representative examples, 'Prospect Avenue' in panel (a) and 'Lower East Side' in panel (b). As shown, there is little overlap between the census tracts and our gentrifying blocks. In panel (a), the gentrifying blocks straddle six different census tracts, and not a single of the latter is fully considered as gentrifying based on our definition. The lack of a one-to-one correspondence between gentrifying blocks and tracts is fairly similar in panel (b). Rapid socio-economic change thus appears highly localized and affects generally only small portions of census tracts.

Figure 5: Examples of gentrifying blocks (blue) vs census tracts (white)

(a) Prospect Avenue (Williamsburg), 1990–2000. (b) Lower East Side, 2000–2010.

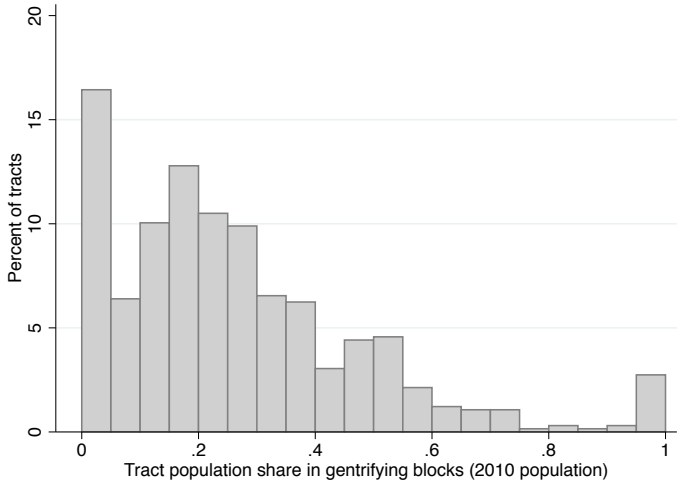
Notes: Gentrifying blocks as identified by our procedure (1990–2000, red; 2000–2010 blue) are overlaid on 2010 census tracts.

The foregoing result holds more broadly and is thus not specific to the two examples we have choosen. For each of the 657 census tracts with at least one gentrifying block in one of the two decades, we compute the share of the population living in the gentrifying blocks. The distribution of these shares is shown in Figure 6. Based on our definition, almost no census tract can be considered as fully gentrifying. The population overlap between gentrifying blocks and census tracts is quite low, 26.5% on average (standard deviation of 22%).

This is again confirmed by anecdotal evidence in the press.

[19] Bayer et al. (2008) is one of the few studies that uses block-level data.

Figure 6: Overlap between 2010 census tracts and gentrifying blocks.



Notes: Histogram of the shares of 2010 tract populations that live in blocks that gentrify between 2000 and 2010.

We run an additional exercise to show that this imperfect overlap is neither due to the criteria we use to define gentrification nor to the fact that we apply them at the block level. We use tract-level data to identify gentrifying tracts in 1990–2000 based on the definition proposed by Couture et al. (2018), i.e., tracts initially below the median real income per capita in their CBSA and with real income per capita growth of at least 50% between 1990 and 2013. When adapted to decennial changes, this amounts to considering tracts with initial real income per capita below the median and growth by 20% between 1990 and 2000. Within a 30 kilometers radius around Wall Street, there are 462 gentrifying tracts out of 3,055 tracts. However, when we consider the blocks within these tracts, nearly 30% of the blocks (1,227 out of 4,184) cannot be considered as gentrifying based on this definition (i.e., they are not below the median in 1990 or experienced a less than 20% increase in their real per capita income over the decade). Conversely, more than 10% of the blocks outside these tracts (5,934 out of 51,561) are gentrifying based on this definition. This means that when all blocks within a tract are assumed to experience the same dynamics, we lose precision in the geographic delineation of the phenomenon we want to analyze.[20]

4 Gentrification and pioneer businesses

Gentrification has been widely studied from a resident's perspective, but rarely from a business' perspective. Notable exceptions include Schuetz (2014), who studies the location de-

[20]Of course, by using the information on income per capita at the block-group level and considering that all the blocks within a block group have the same income per capita, our analysis also partly suffers from measurement error. However, given that there are on average 3.7 block groups within the tracts under study, we believe that our approach represents a valuable improvement in terms of geographic precision of the analysis.

cisions of art galleries in Manhattan and the subsequent redevelopment of neighborhoods; and Couture & Handbury (2017), who propose a list of sectors—restaurants, bars, gyms, and personal services—that may be viewed as consumption amenities for young and educated residents. We adopt a very different perspective. Contrary to previous studies, we uncover the sectors that are associated with gentrification without fixing their list *ex ante*. To do so, we first investigate what businesses are over-represented in 1990 in the blocks that will gentrify between 1990 and 2000. Second, we then ask whether accounting for the presence of these *pioneer businesses* allows us to better predict the future spots of gentrification episodes between 2000 and 2010 in New York, controlling for a large set of covariates. To provide external validity, we then replicate this prediction exercice for another large metropolitan area, Philadelphia.

4.1 Identification of pioneer sectors

We first study whether the locations of businesses in 1990 within some sectors are systematically related to *subsequent* gentrification episodes between 1990 and 2000. To do so, we need to benchmark what the 'expected' location decision of a given establishment should be. We assume that the profit $\pi_{e,b}$ of establishment e in block b linearly depends on a set of location-specific characteristics, x_b, and an unobserved establishment-block specific shock, $\epsilon_{e,b}$:

$$\pi_{e,b} = \alpha + x_b'\beta + \epsilon_{e,b}. \tag{1}$$

It is well known that when the shock $\epsilon_{e,b}$ follows a type-1 extreme value distribution, the probability P_b of choosing block b is of the logit form:

$$P_b = \frac{\mathrm{e}^{\alpha + x_b'\beta}}{\sum_b \mathrm{e}^{\alpha + x_b'\beta}}. \tag{2}$$

Estimating such a logit model for the location choices of more that 600,000 establishments and 60,000 blocks is computationally infeasible. We thus estimate a count model instead. More specifically, we estimate the following negative binomial model:[21]

$$\begin{aligned} n_b^s &= \mathrm{F}(\alpha_0^s + x_b'\alpha^s) + \nu_b^s \\ &= \mathrm{F}(\alpha_0^s + \alpha_1^s lpc\_income_b + \alpha_2^s lpop_b + \alpha_3^s lemp_b + \alpha_5^s lrent_b + \alpha_6^s gentri_b) + \nu_b^s \end{aligned} \tag{3}$$

where n_b^s is the number of establishments in sector s located in block b in 1990; the dummy variable $gentri_b$ identifies blocks that will gentrify between 1990 and 2000; and ν_b^s is a block-sector-specific error term. Our controls x_b' are the logarithms of per capita income, population,

[21] We prefer the negative binomial model to the Poisson count model because it is less sensitive to the excess zero problem—many industries have zero establishments in most blocks—and the overdispersion problem—the dispersion in the number of establishments is much greater than their average number. The list of pioneers obtained using the Poisson estimator is similar to the one we present in the paper.

and employment in 1990 (our proxies for market potential), as well as the logarithm of residential rents in 1990 (our proxy for operating costs). Although residential rents are obviously an imperfect proxy for commercial rents, we do not observe the latter for our blocks. Yet, as shown by Kan et al. (2004), there is on average a positive correlation between residential and commercial rents within blocks, and we think that including the former as controls is better than no controls at all. To smooth our controls around each block, we consider circles of either 250 or 500 meters around the block centroids. Alternatively, we also work with a block-contiguity matrix constructed using GIS. Summing populations and employment, and computing the population-weighted average per capita income in these buffers, provides a better proxy for the market potential of each block.[22] The level of rents is the only variable we consider directly at the block level. We believe that the proxy for operating costs needs to be at the same spatial scale as the dependent variable.

We estimate equation (3) separately for 430 6-digit NAICS industries.[23] Our main coefficients of interest are α_1^s and α_6^s. For each sector, the estimate of the former captures the elasticity of the number of establishments with respect to per capita income; while the estimate of the latter captures the extent to which the sector is overrepresented in 1990 in blocks that will subsequently gentrify between 1990 and 2000.

In what follows, we focus on industries for which $\widehat{\alpha}_6^s > 0$ and significant at the 1% level in at least two of our three specifications (i.e., using the block-contiguity matrix, the 250 meters, and the 500 meters radii). These sectors are overrepresented in areas that will subsequently gentrify, conditional on their market potential and operating costs. Among those sectors, we are further interested in those with a positive income elasticity, i.e., $\widehat{\alpha}_1^s > 0$ and significant at the 1% level. Since gentrifying blocks are, by definition, blocks with income below the median, sectors with a positive income elasticity ($\widehat{\alpha}_1^s > 0$) and a positive gentrification dummy ($\widehat{\alpha}_6^s > 0$) reveal *atypical location choices* that have a high information content. Indeed, these sectors are usually found in wealthy neighborhoods but when they are not found there, they are found in neighborhoods that will change substantially over the next decade.[24] Their presence in poor blocks that are likely to gentrify subsequently thus carries information about the possible causes of gentrification. In the rest of the paper, we refer to establishments in these sectors

[22] Hidalgo & Castañer (2015) find that the contribution of an amenity becomes negligible at about 500 meters, which seems to be the maximum distance that people are willing to move to enjoy it. This distance coincides with the largest radius in our analysis.

[23] There are 792 6-digit NAICS industries with establishments in New York in 1990, but we exclude very small sectors that are active in less than 100 blocks for our estimates to be meaningful. Also, we only keep blocks that have at least one establishment in any sector. Put differently, we remove all purely residential blocks since they may not be suited for any type of activity, e.g., because of zoning laws.

[24] Sectors with $\widehat{\alpha}_1^s > 0$ have a lower probability to locate in below-median income blocks, i.e., it is atypical to find them there. Alternatively, we may view them as picking atypical (poor) blocks, but predominantly those that will end up having higher incomes —which is more in line with the usual location choices of the industry.

as 'pioneer businesses' or simply 'pioneers'. We will show that their presence is important to understand and predict episodes of gentrification, and that the same does not apply to other (non-pioneer) sectors.

Our estimates show that out of the 430 6-digit sectors that we consider in New York in 1990, 28 are overrepresented in below-median income blocks that will subsequently gentrify, and 21 of them are pioneers based on our definition.[25] Table 5 lists them and reports the estimated coefficients of the gentrification dummy across the three specifications (with 250 and 500 meters rings, and contiguous blocks). The list reveals that out of the 21 pioneer sectors that we identify, 18 belong to three 2-digit industries: 7 in "Information and cultural industries" (NAICS 51); 6 in "Arts, entertainment and recreation" (NAICS 71); and 5 in "Professional, scientific and technical services" (NAICS 54). While some of these sectors and professions are frequently mentioned in descriptive studies on gentrification or in newspaper articles, to the best of our knowledge we are the first to provide a list of sectors pioneering gentrification that is derived from a systematic econometric analysis using micro-level data.[26]

Table 5: List of pioneer industries.

NAICS	Name	$\overline{\alpha}_6$	$\widehat{\alpha}_6^{250}$	$\widehat{\alpha}_6^{cont}$	$\widehat{\alpha}_6^{500}$
512131	Motion Picture Theaters (except Drive-Ins)	0.95	0.98	0.99	0.86
711320	Promoters of Performing Arts, Sports, and Similar Events without Facilities	0.88	1.05	0.92	0.66
453920	Art Dealers	0.82	0.92	0.67	0.86
711110	Theater Companies and Dinner Theaters	0.75	0.98	0.79	0.48
512120	Motion Picture and Video Distribution	0.75	0.86	0.84	0.55
541840	Media Representatives	0.73	0.93	0.93	0.31
711510	Independent Artists, Writers, and Performers	0.71	0.83	0.76	0.55
712110	Museums	0.68	0.76	0.75	0.54
512110	Motion Picture and Video Production	0.62	0.71	0.7	0.46
561920	Convention and Trade Show Organizers	0.62	0.82	0.71	0.32
511199	All Other Publishers	0.52	0.54	0.72	0.31
541310	Architectural Services	0.52	0.58	0.61	0.37
711130	Musical Groups and Artists	0.48	0.52	0.57	0.36
541420	Industrial Design Services	0.47	0.58	0.52	0.31
541430	Graphic Design Services	0.43	0.44	0.54	0.31
813990	Other Similar Organizations (except Business, Professional, Labor, and Political Organizations)	0.43	0.41	0.56	0.32
511120	Periodical Publishers	0.43	0.54	0.57	0.17
711410	Agents and Managers for Artists, Athletes, Entertainers, and Other Public Figures	0.42	0.49	0.53	0.23
541922	Commercial Photography	0.41	0.41	0.53	0.28
518210	Data Processing, Hosting, and Related Services	0.39	0.44	0.51	0.21
511130	Book Publishers	0.29	0.41	0.39	0.08

Notes: Estimates of $\widehat{\alpha}_6$ conditional on $\widehat{\alpha}_1 > 0$. We define $\overline{\alpha}_6$ as $\overline{\alpha}_6 = (\widehat{\alpha}_6^{250} + \widehat{\alpha}_6^{500} + \widehat{\alpha}_6^{cont})/3$, i.e., the average of the coefficients obtained on the gentrification dummy across the three specifications—250 meters buffer, 500 meters buffer, contiguous blocks—that we estimate.

[25]Hence, 75% of the overrepresented sectors in soon-to-gentrify blocks tend to usually locate in wealthier blocks. The 7 remaining industries are mainly light manufacturing or related to the music industry ("Prerecorded Compact Disc (except Software), Tape, and Record Reproducing" and "Sound Recording Studios").

[26]One such well-known example is art galleries or "Art dealers" (NAICS 453920). See Schuetz (2014) and Easterly et al. (2018) for case studies; and `https://urbanize.la/post/gentrification-and-arrival-art-galleries-boyle-heights-there-correlation` for a popular press account.

4.2 Identification of pioneer sectors: Sensitivity tests

To assess the robustness of our list of pioneers, we run two exercises. First, we reestimate equation (3) using alternative definitions of gentrification. Following the tract-level definition in Couture et al. (2018), we consider as gentrifying the blocks with initial per capita income below the median and subsequent growth of at least 20% over the following decade. According to this definition, 8,643 blocks are gentrifying—instead of 1,381 based on our benchmark definition. The list of pioneer sectors obtained using this alternative definition contains only 5 sectors that are all in our benchmark list ("Art dealers", "Motion picture and video production", "Architectural services", "Theater companies and dinner theaters" and "Independent artists"). We thus lose a lot of industries from the initial list using these alternative criteria to define gentrification.

Second, we consider a change in the geographic units used for the analysis. To this end, we aggregate our data to the tract level and use the same definition of gentrification as in Couture et al. (2018). We then obtain a list of 32 pioneer sectors, with 22 industries from the information, entertainment, and creative business services industries. However, the list now also includes some industries from the finance and real estate sectors and from the healthcare sector ("Offices of Physicians, Mental Health Specialists"). At the same time, we lose several activities among which "Motion picture and video distribution", "Media representatives", or "Musical groups and artists".

The main message from these robustness checks is that the businesses from the information, cultural, art, entertainment and creative business services industries are robustly associated with gentrification, whatever the exact criteria and the spatial scale used to define gentrification. However, the length and composition of the list change with the definition of gentrification. Using laxer criteria reduces drastically the list; whereas using a more aggregate spatial scale makes it less focused on information, cultural, art, entertainment and creative business services industries (90% in the benchmark list against 70% in the tract-level list).

4.3 Are pioneers heralds of gentrification?

As explained before, our pioneers are initially overrepresented in places that will subsequently gentrify. Yet, this gentrification process could be entirely driven by other initial characteristics of the blocks. In particular, demographic characteristics of the population, crime, the age of the housing stock, and various amenities could drive both the initial location of pioneer businesses and the subsequent arrival of more affluent and educated residents. We thus now investigate whether the initial presence of pioneers significantly predicts subsequent gentrification, controlling for a large number of initial block-level characteristics. For this prediction exercise to have external validity, we rely on a different decade, or a different city, or both than those used for the identification of the pioneer sectors. Namely, we estimate the determinants of gentrifi-

cation in New York and Philadelphia during the decade 2000–2010, using initial characteristics of the blocks as regressors. The explanatory variables include a *block-level measure of exposure to pioneer businesses* in 2000, which is constructed using the list of pioneer sectors estimated for New York in 1990–2000.

4.3.1 Methodology

We estimate the probability that a block gentrifies, taking into account five types of initial characteristics. First, we control for socio-economic variables such as population size, per capita income, the share of educated (defined as those with at least a college degree), and the block's racial composition. Second, we control for housing characteristics using rents and the median age of buildings. Third, we include several proxies for 'amenities': distance to parks, distance to public transportation, a dummy for waterfront blocks, the number of landmarks, and—for the subsample of blocks in the five boroughs of New York City—the number of property crimes and of violent crimes. Fourth, we account for spatial contagion in the gentrification process (see Figure 2) by controlling for the distance to the closest block that gentrified over the preceding decade. Finally, our main variable of interest is the block-level measure of exposure to pioneer businesses, which may be a possible determinant of gentrification. The latter is directly constructed using the identification of pioneer sectors described in Section 4.1. For each pioneer sector $i \in \mathcal{P}$, where $\mathcal{P}$ is the set of pioneer industries, we compute the average $\overline{\alpha}_6^i$ of the coefficients obtained on the gentrification dummy (using 250 and 500 meters radii, as well as contiguous blocks). We then compute the weighted sum of pioneer businesses in the block, using these average coefficients as weights. Formally, we define the exposure $expo_b$ of block b to pioneers as:

$$expo_b \equiv \ln\left(1 + \sum_{i \in \mathcal{P}} \overline{\alpha}_6^i \times estab_b^i\right), \tag{4}$$

where $\overline{\alpha}_6^i = (\widehat{\alpha}_6^{250} + \widehat{\alpha}_6^{500} + \widehat{\alpha}_6^{cont})/3$ is the average coefficient on the gentrification dummy for sector i, and $estab_b^i$ is the number of pioneer establishments in sector i located in block b. Observe that our measure (4) assigns more weight to sectors that display more atypical location patterns—as captured by $\overline{\alpha}_6^i$—since those carry more information.

We restrict our analysis to blocks that can potentially gentrify between 2000 and 2010, i.e., blocks with per capita income below the median and with at least 8 residents in 2000. Our dependent variable is a dummy $\mathbb{1}_b^{gentri}$, equal to one if the block gentrifies during 2000–2010 and zero otherwise. All explanatory variables are measured in 2000 and computed within 250 meter rings around the centroid of each block, except for the waterfront dummy and the distance variables. We estimate

$$\mathbb{1}_b^{gentri} = \alpha + x_b'\beta + expo_b\gamma + \epsilon_b \tag{5}$$

using a linear probability model as it is more amenable to instrumental variables estimations. Finally, we correct for the spatial correlation of the standard errors in (5) in a radius of 500 meters around each block using the method proposed by Conley (1999) and the Stata package developed by Fabrizio et al. (2018), both for OLS and IV.

4.3.2 Predictions for New York and Philadelphia, 2000–2010

Table 6 contains our results for 2000–2010. Columns (1)–(4) show the results for New York, while columns (5)–(6) report the results for Philadelphia.

Baseline results. Starting with New York, the following picture emerges from all specifications that we estimate. Among the blocks with below-median per capita income, those that gentrify are initially less populated, more affluent, and have a higher share of highly educated residents. Hence, gentrifying blocks are not found among the poorest and most disadvantaged ones. 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 these blocks tend to be older and less expensive as measured by rents. This is in accord with the literature that has emphasized the importance of neighborhood housing cycles induced by the deterioration and subsequent rebuilding of the housing stock (Rosenthal 2008, Brueckner & Rosenthal 2009). Futhermore, there is contagion: all else equal, the closer a block is to a block that gentrified in the previous decade, the higher the probability that it also gentrifies subsequently. As to amenities, gentrifying blocks tend to be closer to the waterfront, subway stations, and parks, though the coefficients are not always significant. Finally, regarding our key variable—the blocks' exposure to pioneers—the results are very clear. Column (1) shows that the exposure to pioneers in a 250 meters radius around a block in 2000 is positively correlated with the probability that this block will gentrify between 2000 and 2010. We also control for the total number of (non-pioneer) establishments in a 250 meters radius around the block. Strikingly, our results show that only pioneer businesses—which display atypical location patterns—are positively correlated with subsequent gentrification episodes. Indeed, the coefficient on the number of other (non-pioneer) establishments is never significantly positive.

There is an abundant literature on crime and gentrification (see, e.g., O'Sullivan 2005). Some authors even argue that falling crime is the major determinant of subsequent gentrification (Ellen et al. 2017). In column (2), we reproduce the analysis for the blocks in the five boroughs of New York City, for which we can also control for the initial level of crime. Property crimes (robbery and burglary) seem to have a significant—though ambiguous—relationship with gentrification; and high initial levels of violent crimes (murder and rape) are associated with a lower probability of subsequent gentrification, though these coefficients are not precisely

Table 6: Determinants of gentrification in New York and Philadelphia, 2000–2010.

Dependent variable	Dummy equal to one for blocks that gentrify during 2000–2010, $\mathbb{1}_b^{gentri}$.					
	New York				**Philadelphia**	
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to pioneers	0.058^a	0.054^a	0.064^a	0.063^a	0.074^a	0.064^b
	(0.009)	(0.010)	(0.013)	(0.016)	(0.017)	(0.026)
Number of establishments (log)	-0.009^b	-0.003	-0.013^b	-0.008	0.001	-0.002
	(0.004)	(0.006)	(0.005)	(0.007)	(0.006)	(0.008)
Per capita income (log)	0.006	0.050^c	0.003	0.047^c	-0.024	-0.025
	(0.019)	(0.026)	(0.019)	(0.026)	(0.031)	(0.031)
Population (log)	-0.009^c	-0.016^b	-0.008^c	-0.015^b	-0.020^b	-0.017^b
	(0.005)	(0.007)	(0.005)	(0.007)	(0.008)	(0.008)
Share college educated	0.318^a	0.218^b	0.314^a	0.210^b	0.356^a	0.376^a
	(0.078)	(0.100)	(0.080)	(0.101)	(0.131)	(0.133)
Share African-American	-0.006	-0.028	-0.007	-0.030	-0.092^a	-0.093^a
	(0.014)	(0.021)	(0.014)	(0.021)	(0.022)	(0.022)
Share Asian	-0.102^b	-0.147^a	-0.101^b	-0.147^a	-0.122	-0.107
	(0.040)	(0.043)	(0.040)	(0.043)	(0.149)	(0.152)
Share other minority residents	-0.011	-0.021	-0.010	-0.021	-0.094^c	-0.095^c
	(0.043)	(0.052)	(0.043)	(0.052)	(0.051)	(0.051)
Rent (log)	-0.025^c	-0.026	-0.025^c	-0.025	-0.005	-0.007
	(0.015)	(0.019)	(0.015)	(0.019)	(0.026)	(0.026)
Median age of buildings (log)	0.001^a	0.002^a	0.001^a	0.002^a	0.003^a	0.003^a
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Dist. to closest gentrifying block 1990–2000 (log)	-0.011^b	-0.009	-0.011^b	-0.008	-0.029^a	-0.029^a
	(0.004)	(0.006)	(0.004)	(0.006)	(0.009)	(0.010)
Less than 200m from waterfront	0.013	0.030	0.013	0.029	0.007	0.006
	(0.013)	(0.024)	(0.013)	(0.024)	(0.019)	(0.019)
Distance to closest park (log)	-0.004^c	-0.008^a	-0.004^c	-0.008^a	0.003	0.004
	(0.002)	(0.003)	(0.002)	(0.003)	(0.005)	(0.005)
Distance to subway (log)	-0.009^a	-0.003	-0.009^a	-0.003	-0.000	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
# of main landmarks (log)	0.001	0.016	0.001	0.015	0.013	0.014
	(0.009)	(0.013)	(0.009)	(0.013)	(0.016)	(0.015)
Robbery		0.033^a		0.032^a		
		(0.006)		(0.006)		
Burglary		-0.018^a		-0.017^a		
		(0.006)		(0.006)		
Murder		-0.227		-0.218		
		(0.147)		(0.145)		
Rape		-0.102		-0.103		
		(0.105)		(0.105)		
# of observations	34,164	19,868	34,164	19,868	18,144	18,144
Sample	New York	NYC	New York	NYC	Philadelphia	Philadelphia
Specification	OLS	OLS	IV	IV	OLS	IV

Notes: The sample is composed of blocks with per capita income below the median in the city in 2000, and with at least 8 residents. The measure of exposure to pioneers is given by (4). All explanatory variables are measured in 2000 and are computed using 250 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 parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

measured.[27] Observe that our estimates of the other covariates, in particular our measure of exposure to pioneers, are barely affected by the inclusion of the crime measures.

Instrumental variables estimates. The presence of pioneers in 2000 might be correlated with unobserved block-level shocks that are the actual drivers of gentrification. We propose an IV strategy—based on a Bartik-type instrument—to address this potential endogeneity concern. The instrument is designed as follows. First, we consider for each pioneer sector the stock of businesses within a 250 meters radius around the block in 1990. Second, we apply to this stock the NAICS 6-digit growth rate in the number of plants in this sector observed in another similar north-eastern city—Boston—between 1990 and 2000. We then use these values to build a predicted measure of exposure to pioneers based on equation (4). Our instrument is strongly and positively correlated with the instrumented variable (correlation of 0.78, first stage R^2 of 0.67), thus showing its relevance. We use a similar instrument for the count of establishments in the other (non-pioneer) sectors.

Bartik-type instruments have been subject to growing criticism in recent years. However, we believe that the use of such an instrument is safe in our context. Indeed, it combines block-level economic activity in New York in 1990 with sectoral evolutions in Boston between 1990 and 2000. Our dependent variable is based on block-level demographic changes in New York between 2000 and 2010. Controlling for all covariates that we include in our regressions, it seems reasonable to assume that sectoral growth rates in Boston between 1990 and 2000 are disconnected from unobserved shocks that affect block-level demographic changes in New York and Philadelphia in the subsequent period. Regarding the stock of establishments around the block in 1990, the identifying assumption is that the entrepreneurs running these establishments in 1990 were not aware of possible block-level shocks actually triggering gentrification over the period 2000–2010. This seems all the more reasonable as most pioneer establishments in 2000 were not active in 1990.[28]

Columns (3) and (4) in Table 6 corroborate our previous findings: the presence of pioneer businesses positively affects the probability that a block gentrifies over the subsequent decade. Note that all coefficients in our IV estimations remain fairly similar to their OLS counterparts, thus suggesting that any potential endogeneity bias may be small.

Results for Philadelphia. Columns (5) and (6) in Table 6 report estimation results for another metropolitan area in the north-east, Philadelphia. To run these regressions, we have built a sta-

[27] The HAC correction that we apply is very demanding. The coefficients on murder and rape are highly significant when computed using robust standard errors without that correction.

[28] Among the establishments that were present in the blocks in 2000, the median year of their first appearance in our dataset is 1993 for non-pioneer industries, both in the blocks that will gentrify between 2000 and 2010 and in the other blocks. For pioneer industries, this median year is 1995 in the blocks that will gentrify between 2000 and 2010, but 1993 in the other blocks.

ble geography for Philadelphia and replicated our methodology to identify gentrifying blocks over the period 2000–2010. We then examine the determinants of gentrification, including a measure of exposure to pioneers computed using establishment data for Philadelphia together with the list of pioneer industries identified in New York over the decade 1990 to 2000. We do not have the geocoded data to control for crime in Philadelphia. As Table 6 shows, the results are very similar to those obtained in the previous columns: blocks more exposed to pioneers in 2000 are more likely to gentrify over the subsequent decade.

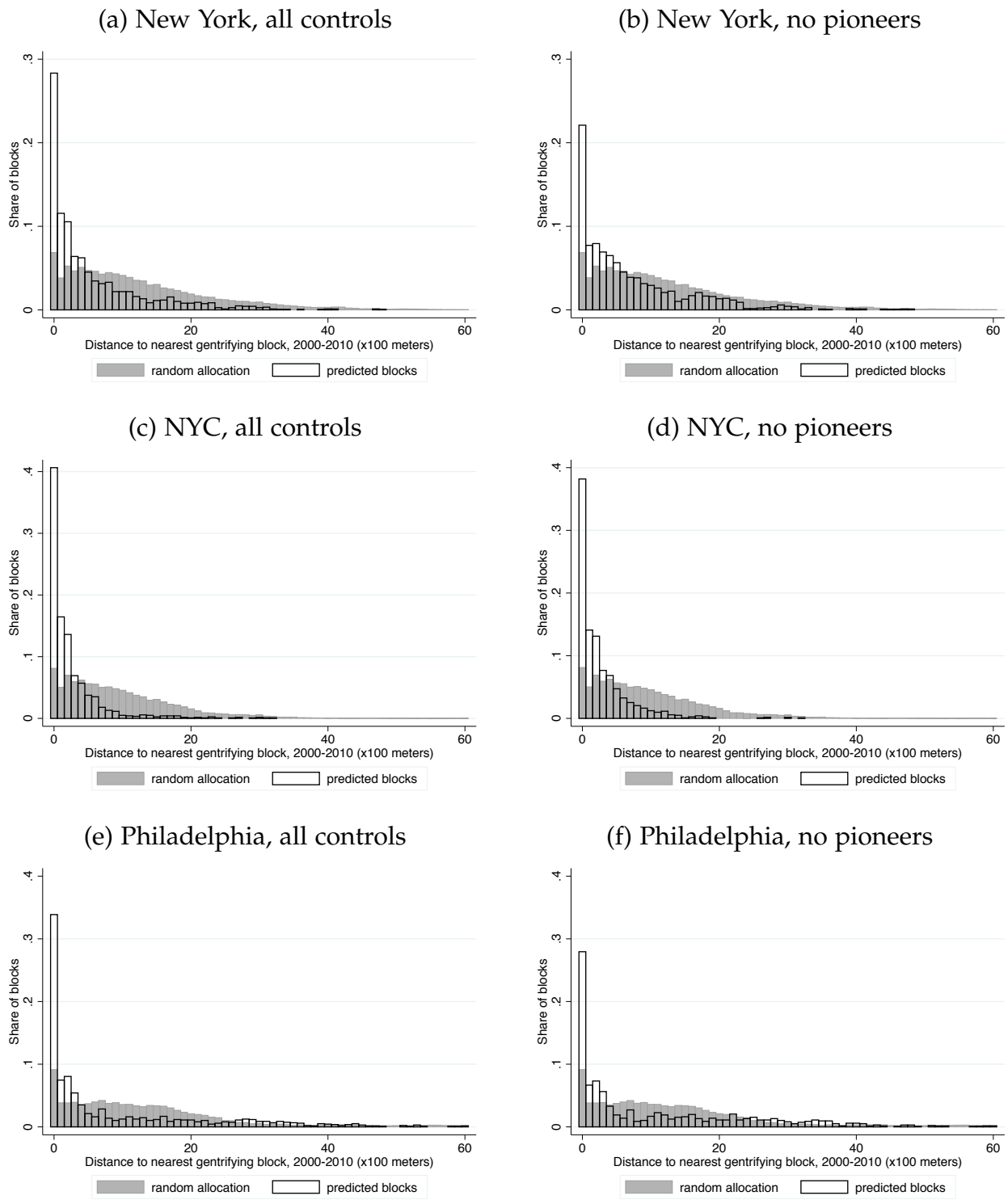
Geographic scale. As a final robustness check, we verify that our results are not sensitive to the geographic scale used to construct the explanatory variables. Computing them using 500 meters rings or contiguous blocks—instead of 250 meters rings—yields qualitatively very similar results (see Tables 11 and 12 in Appendix D).

4.4 Quantitative importance of pioneers

Pioneers help explain where gentrification takes place. We want to assess how our empirical model performs in predicting gentrification spots. To this end, we compute—for each eligible block in New York and Philadelphia, i.e., blocks with below-median per capita income and at least 8 residents—the predicted probability that it gentrifies between 2000 and 2010 based on our regression coefficients. We then rank the blocks based on this predicted probability and consider the blocks with the highest predicted probability up to rank N, where N is the number of blocks that actually gentrified in the MSA ($N = 1,866$ in New York, $N = 1,228$ in NYC, and $N = 1,358$ in Philadelphia between 2000 and 2010). We can then compute the share of those N blocks which effectively gentrified, or which are within a given distance from a block that actually gentrified. The first column of Table 7 shows that we correctly predict block-level gentrification in 23%, 32%, and 30% of cases in New York, NYC, and Philadelphia, respectively. While these figures might seem low, they reflect the high level of idiosyncrasy in the gentrification process highlighted in Hwang & Lin (2016) and Easterly et al. (2018). Moreover, adopting a less stringent criterion to assess the predictive power of our model, we observe that at least 50% of the blocks with the highest probability to gentrify in New York and Philadelphia are less than 300 meters away from a block that actually gentrified between 2000 and 2010 (and nearly 60% less than 500 meters away).

The predictive power of the model is *much higher than that associated with a coin toss*. We report in the left panels of Figure 7 the spatial distribution of the blocks that are predicted to gentrify based on both our empirical model and a random allocation for New York and Philadelphia. In the two samples, our empirical model increases dramatically the share of predicted blocks that are located in a 500 meter radius from a block that actually gentrified relative to a random model.

Figure 7: Distance of predicted from actual blocks.



Notes: Predictions based on equation (5). The random allocation is based on the average of 1,000 draws of N blocks.

Table 7: Quantitative importance of pioneers and contagion.

Distance (meters)	All controls	No pioneers	No contagion	No pioneers and no contagion
	(1)	(2)	(3)	(4)
	New York (30km around Wall Street)			
Exact	0.23	0.19	0.21	0.19
0-100	0.28	0.22	0.27	0.22
100-200	0.40	0.30	0.38	0.30
200-300	0.50	0.38	0.49	0.38
300-400	0.57	0.45	0.56	0.45
400-500	0.63	0.51	0.62	0.52
	NYC			
Exact	0.32	0.31	0.32	0.30
0-100	0.41	0.38	0.41	0.37
100-200	0.57	0.52	0.57	0.50
200-300	0.71	0.65	0.70	0.63
300-400	0.78	0.73	0.77	0.71
400-500	0.83	0.80	0.83	0.78
	Philadelphia (30km around the CBD)			
Exact	0.30	0.25	0.24	0.17
0-100	0.34	0.28	0.27	0.18
100-200	0.41	0.35	0.34	0.22
200-300	0.49	0.42	0.42	0.28
300-400	0.55	0.48	0.49	0.34
400-500	0.58	0.51	0.53	0.37

Notes: We report the share of blocks with the highest predicted probability to gentrify from the closest blocks that actually gentrify by one hundred meter distance bins. 'Exact' refers to the share of effectively gentrifying blocks that are predicted to gentrify (i.e., a distance of 0 meters).

Pioneers have significant predictive power. Our foregoing results show that the model significantly improves our ability to predict the geography of gentrification as compared to a random allocation of the phenomenon. We now further isolate the quantitative importance of pioneers in our model. To gauge their contribution, we repeat the same exercise as before but without using information on exposure to pioneers. The results are presented in the right panels of Figure 7 and the associated figures are reported in column (2) of Table 7. In the two samples, including our measure of exposure to pioneers significantly improves the predictive power of the model. For New York, the share of correct block-level predictions of gentrification drops by 4 percentage points (p.p.) if we do not use information on pioneers. Including the exposure to pioneers improves the share of correct predictions within 100 meters by 6 p.p., within 200 meters by 10 p.p., and within 300 meters by 12 p.p. For Philadelphia, pioneers contribute 7 p.p. to the increase in predictions in a 300 meters radius. These figures are overall rather big since they represent an increase in the number of "correct" predictions within a 100m radius around gentrifying blocks by 21% in New York and 18% in Philadelphia.

Pioneers matter at least as much as contagion. We replicate our exercise without the spatial contagion variable, which allows us to gauge the relative importance of our exposure to pio-

neers variable for the predictive power of the model. Columns (3) and (4) of Table 7 summarize our results. For New York, contagion does not seem to have a large effect as compared to pioneers. Things are different for Philadelphia, with pioneers and contagion contributing both 7 p.p. of correct predictions within a 300 meters radius. Our quantification exercise thus suggests that exposure to pioneer businesses contributes at least as much as spatial contagion to the prediction of future spots of gentrification. We next explore why the presence of pioneers may have such an influence on local neighborhood dynamics.

5 Why do pioneers herald gentrification?

Thus far we have: (i) identified a list of sectors whose presence systematically precedes gentrification episodes, i.e., sectors that are overrepresented in soon-to-gentrify neighborhoods conditional on block-level characteristics; and (ii) shown that these sectors contribute substantially to improve the prediction of where gentrifying areas are located, on top of other socio-economic variables. However, until now we have not looked at the mechanisms through which these pioneer businesses may affect gentrification. In this section, we strive to identify and to better understand some of these channels.

We consider several mutually non-exclusive explanations for the role of pioneers in the gentrification process. First, we study the link between residents and establishments by looking at the type of workers employed in pioneer sectors. We will show that pioneers tend to employ people whose socio-economic characteristics are traditionally associated with gentrifying neighborhoods. Second, we investigate the influence of pioneers on the attractiveness of a neighborhood. We find that pioneers are associated with rising housing prices, yet that they cannot themselves be viewed as consumption amenities. Last, we study the relationship between pioneers and consumption amenities. We find that pioneers precede the later arrival of consumption amenities in a neighborhood.

5.1 Workers in pioneer industries are different

Pioneer industries may employ workers with specific characteristics that can play a role in the subsequent gentrification process. To investigate this possibility, we first examine the socio-demographic profile of workers employed in pioneer industries. We use IPUMS microdata for the years 2000 and 2010 to compute a set of worker characteristics in the New York MA, distinguishing pioneer from non-pioneer industries.[29]

Figure 8 reveals systematic differences in the characteristics of workers employed by the two types of industries. First, as shown by panels (a) and (b), pioneer sectors employ younger and more educated workers. However, wages in pioneer sectors are not markedly different from

[29]Results using data for the entire U.S. are very similar and relegated to Figure 11 in Appendix D.

wages in the other sectors. As shown by panels (d) to (f), these workers are more often single or in a power couple (i.e., couples in which both members are college educated; Costa & Kahn 2000), 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 by bicycle or by foot more often than workers in other industries.

The big picture that emerges from the foregoing facts is that workers in pioneer industries tend to live closer to their workplace, which suggests the existence of a connection between the presence of pioneer establishments and the socio-demographic composition of the neighborhood where they are located. In addition, all of the aforementioned characteristics (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 & Hartley 2016, Couture & Handbury 2017); (ii) these millennials work with more flexible employment relationships than the older 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 young educated 'millennials' seem to 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 tight connection in the data between the mix of businesses in a given area and the subsequent evolution of this area in terms of residents. This holds especially true if pioneer residents act as magnets who attract other affluent people or valuable consumption amenities. We now turn to those two points.

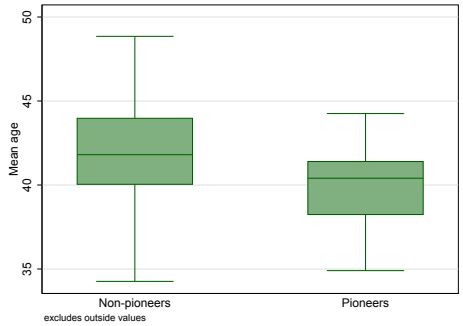
5.2 Pioneers attract affluent residents

The presence of pioneer establishments in a relatively poor (and inexpensive) neighborhood may attract wealthy people who are not directly active in those businesses. In the presence of uncertainty, for instance, the location of pioneers may provide information on a neighborhood's prospects for potential investors or other businesses. In that case, the presence of pioneers might constitute a signal that the neighborhood will experience rapid upwards mobility in the near future—that it will become the "next hot neighborhood"—which may trigger rapid changes (see, e.g., Caplin & Leahy 1998, for a model of herding and businesses). Pioneer businesses could also be amenities that are valued by wealthy people. If the latter are willing to live close to these establishments, their presence in a neighborhood will induce a new influx of affluent residents in the area, which should be capitalized into higher housing values.

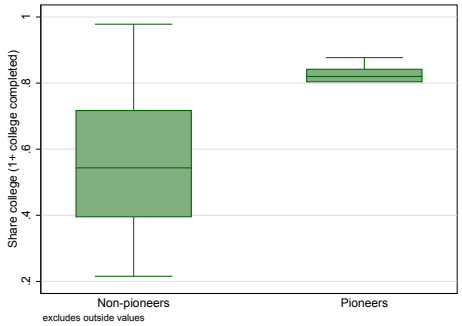
There are no direct measures of the signal provided by pioneers or the amenities they offer to residents. However, we can look at measures of residential attractiveness to indirectly test these two mechanisms. In panels (a) and (b) of Table 8, we estimate simple hedonic

Figure 8: Worker characteristics in pioneer and non-pioneer industries in New York.

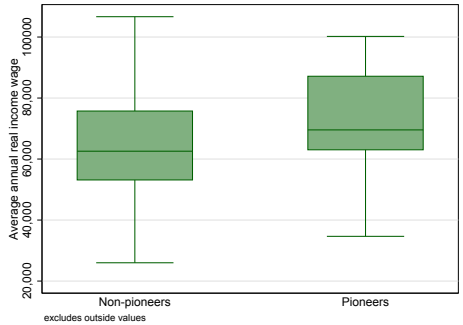
(a) Pioneers are younger.



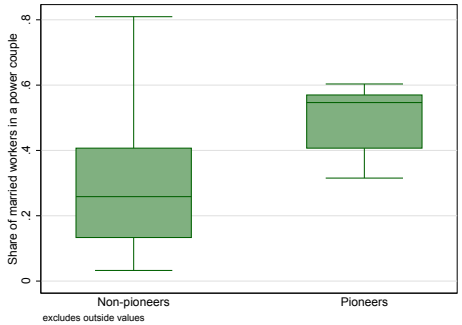
(b) Pioneers are more educated.



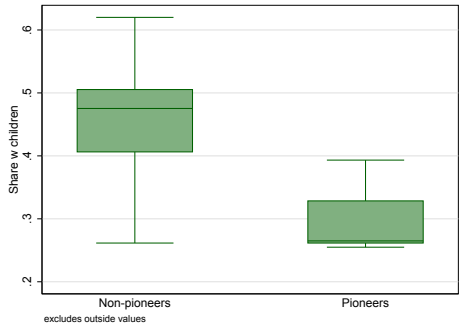
(c) Pioneers are not much wealthier.



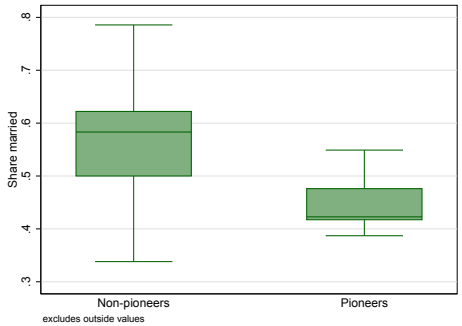
(d) Pioneers are in 'power couples'.



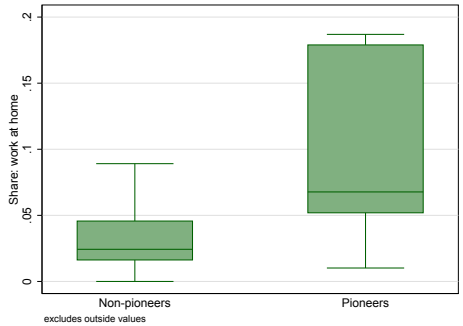
(e) Pioneers have fewer children.



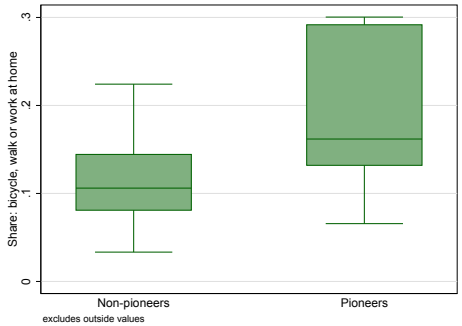
(f) Pioneers are more often single.



(g) Pioneers work more at home.



(h) Pioneers commute shorter distances.



Notes: Our computations using IPUMS data for the years 2000 and 2010.

Table 8: Pioneers and housing prices in 1990.

Dependent variable	Block-level housing prices in 1990.			
	(1)	(2)	(3)	(4)
Panel (a) Dependent variable: median rents in 1990				
Number of pioneers (log)	0.008^a	0.055^a	-0.006^a	-0.014^a
	(0.002)	(0.002)	(0.002)	(0.002)
Observations	55,078	55,078	55,075	55,062
R^2	0.000	0.128	0.511	0.554
Panel (b) Dependent variable: median housing values in 1990				
Number of pioneers (log)	0.116^a	0.133^a	-0.008^b	-0.022^a
	(0.003)	(0.003)	(0.003)	(0.003)
Observations	53,189	53,189	53,186	53,181
R^2	0.047	0.062	0.552	0.611
Physical amenities	No	Yes	Yes	Yes
Demographic characteristics	No	No	Yes	Yes
Per capita income	No	No	No	Yes

Notes: The sample is composed of all non-gentrifying blocks in 1990. 'Number of pioneers' is the count of pioneer establishments within 250 meters around the block. All explanatory variables are measured in 1990 and are computed using 250 meter rings around each block (except the distance to subway, to parks and to closest gentrifying block variables, as well as the waterfront dummy). See Table 6 for the complete list of explanatory variables. Huber-White robust standard errors in parenthesis. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

regressions using the median rent and the median housing value in each block in 1990 as the dependent variable, respectively. We restrict our sample to non-gentrifying blocks, to avoid that the specific features of gentrifying blocks affect our results, and we use the same variables than those we use in Table 6.

We first verify that the presence of pioneers is associated with higher residential attractiveness, even outside gentrifying areas. If this were not the case, it would be difficult to argue in favor of either of these two channels. Column (1) of Table 8 shows that the exposure to pioneers is associated with higher housing prices. This result is consistent with a signalling effect: since pioneer establishments usually tend to locate in neighborhoods with higher rents or housing values, seeing such establishments in poorer or more deprived areas might signal a positive future evolution of the neighborhood. Since this positive coefficient may easily capture the influence of other variables, we include in column (2) a set of 'physical' characteristics—proximity to parks, subways, and waterfront. The positive coefficient survives the inclusion of these characteristics.

The estimated coefficient turns, however, negative in column (3) once we control for the presence of other (non-pioneers) establishments, as well as the demographic characteristics of the area (population, education, and racial composition). Finally, we control for per capita

income in column (4). Our results show that once block-level characteristics and income are taken into account the presence of pioneer establishments is not significantly associated with higher prices—indeed, the opposite holds true. Compensating differentials theory thus suggests that pioneer businesses do not constitute valuable amenities for residents (Roback 1982). This is not surprising as the vast majority of our pioneer industries do not provide goods or services that can be directly consumed by residents.[30]

5.3 Pioneers attract consumption amenities

Recent results by Couture & Handbury (2017) suggest that the revival of American inner-cities has been triggered by a shift in the preferences of highly educated workers towards certain non-tradable consumption amenities—such as restaurants, bars, and gyms. These services are not in our list of pioneers. However, it might be that the presence of pioneers in an area—and their associated workers—attract these amenities which, in turns, leads to a change in the mix of residents. For example, the presence of pioneers may generate a local market for consumption amenities. If those consumption amenities are also valued by wealthy people, the latter may choose to locate in these areas, thereby changing the composition of the neighborhood.

To test this mechanism, we investigate whether the presence of pioneers (within 250 meters around the block) in 1990 is associated with the net arrival of consumption amenities in the block during the next decade—controlling for the initial level of amenities. The results are summarized in Table 9. In columns (1) and (2), we use a sample that includes all blocks in New York—except the ones that gentrified during the period—while we include only the 'poor' blocks in columns (3) and (4), i.e., the blocks with per capita income in 1990 below the median of the metropolitan area. We adopt the classification of Couture & Handbury (2017) and estimate the model for three different types of amenities: non-tradable services (restaurants, bars, personal services, and sports), stores, and activities (museums, galleries, libraries, parks, and golfs).

Our results show that the early presence of pioneers in a block is significantly associated with a subsequent net influx of consumption amenities. The results in columns (2) and (4) show that this positive association holds even when we control for the initial characteristics of the block. The only exception is for stores in relatively poor areas, where the presence of pioneers has no effect on subsequent net entry of establishments. Overall, we find robust

[30]We explore this result further by analyzing the evolution of pioneers' performance in gentrified neighborhoods. If pioneers are amenities consumed by local residents, one would expect an increase in their performance as demand for their products and services rises during a gentrification episode. We thus investigate whether pioneer businesses grow faster, are more likely to survive and see their credit scores improve compared to the other businesses in gentrifying areas. The results are relegated to Table 13 in Appendix D. They show that there are no statistical differences between establishments in pioneer and in non-pioneer industries. Hence, pioneers do not seem to be amenities that are particularly valued by new affluent residents.

evidence that the presence of pioneers is associated with the arrival of consumption amenities in a block. The latter might then trigger the arrival of affluent residents.

Table 9: Exposure to pioneers and net changes in consumption amenities, 1990–2000.

Dependent variable	Δ log (# of consumption amenities)			
	(1)	(2)	(3)	(4)
Panel (a): Non-tradable services				
Initial # of pioneers (log)	0.059^a	0.013^a	0.077^a	0.011^b
	(0.002)	(0.003)	(0.004)	(0.004)
Panel (b): Stores				
Initial # of pioneers (log)	0.047^a	0.020^a	0.047^a	0.006
	(0.002)	(0.003)	(0.004)	(0.004)
Panel (c): Activities				
Initial # of pioneers (log)	0.007^a	0.005^a	0.004^a	0.003^b
	(0.001)	(0.001)	(0.001)	(0.001)
Sample	Non-gentrifying blocks		Poor blocks	
Block-level controls	No	Yes	No	Yes
Observations	62,418	55,010	32,578	32,399

Notes: The sample is composed of all non-gentrifying blocks in 1990 in columns (1) and (2); and of blocks with income per capita below the median in the city in 1990 in columns (3) and (4). Consumption amenities are defined as in Couture & Handbury (2017). 'Initial # of pioneers' is the count of pioneer establishments within 250 meters around the block at the beginning of the period. All explanatory variables are measured in 1990 and are computed using 250 meter rings around each block (except the distance to subway, to parks and to closest gentrifying block variables, as well as the waterfront dummy). See Table 6 for the complete list of explanatory variables. Huber-White robust standard errors in parenthesis. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

6 Conclusion

Gentrification is an important phenomenon that attracts substantial media and public policy attention. Yet, it is fair to say that it is still poorly understood. Where does it occur? Who are the potential "gentrifiers"? What are the possible roles of businesses in that process? To make progress on these questions, we have built a time-consistent block-level dataset with socio-economic characteristics of residents and information on businesses in New York from 1990 to 2010. Using these geographically fine-grained data, we have first identified gentrifying areas. The picture that emerges is that gentrification is both highly localized and spread across the metropolitan area. Thus, an analysis beyond the central city and at a small geographic scale seems required to capture the full extent of that phenomenon.

Second, we have identified pioneers, i.e., industries in which establishments tend to make

atypical location decisions. Our endogenously identified list of pioneers subsumes several cultural, recreational, and creative industries. In particular, cultural and artistic activities feature prominently in it, thus providing solid quantitative evidence for the numerous stories and case studies about 'artists and gentrification'.

Third, we have shown that a block's exposure to pioneers has a quantitatively sizable effect in predicting future gentrification. Controlling for a large set of initial characteristics of a block, the block's exposure to pioneers is positively related to the probability that it will gentrify during the next decade. This effect survives a battery of tests and is robust to IV estimation and controlling for spatial correlation. It is also quantitatively sizable: it increases the share of blocks that are correctly predicted to gentrify by more than 20% within 100 meters distance from blocks effectively gentrifying in New York, and by about 18% in Philadelphia. The magnitude of the effect is comparable to that of spatial contagion.

Last, we have provided suggestive evidence on the mechanisms by which pioneers may fuel gentrification. The likely channels are through the types of workers—young and educated, without children, and in power couples—they hire, the signal they produce regarding the future prospects of the neighborhood, and their effect on the subsequent arrival of consumption amenities valued by affluent and educated residents. Pioneer businesses seem to play an important role in the gentrification process, but further research is required to disentangle and to quantitatively assess the contribution of the various possible mechanisms to this phenomenon.

References

Aguiar, M., Bils, M., Charles, K. K. & Hurst, E. (2017), 'Leisure luxuries and the labor supply of young men', NBER Working Papers 23552, National Bureau of Economic Research, Inc.

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

Baum-Snow, N. (2014), 'Urban transport expansions, employment decentralization, and the spatial scope of agglomeration economies', *Processed, Brown University* .

Baum-Snow, N. & Hartley, D. (2016), 'Accounting for central neighborhood change, 1980-2010', Working Paper Series WP-2016-9, Federal Reserve Bank of Chicago.

Bayer, P., Ross, S. L. & Topa, G. (2008), 'Place of work and place of residence: Informal hiring networks and labor market outcomes', *Journal of Political Economy* **116**(6), 1150–1196.

Bernard, A. B., van Beveren, I. & Vandenbussche, H. (2012), 'Concording EU trade and production data over time', CEPR Discussion Papers 9254, C.E.P.R. Discussion Papers.

Bostic, R. W. & Martin, R. W. (2003), 'Blackhome-owners as a gentrifying force? Neighbourhood dynamics in the context of minority home-ownership', *Urban Studies* **40**(12), 2427–2449.

Brown, A., Blumenberg, E., Taylor, B., Ralph, K. & Voulgaris, C. (2016), 'A taste for transit? Analyzing public transit use trends among youth', *Journal of Public Transportation* **19**(1), 49–67.

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

Brummet, Q. & Reed, D. (2018), 'Gentrification and the well-being of original neighborhood residents: Evidence from longitudinal census microdata', Mimeo.

Burchfield, M., Overman, H. G., Puga, D. & Turner, M. A. (2006), 'Causes of sprawl: A portrait from space', *The Quarterly Journal of Economics* **121**(2), 587–633.

Caplin, A. & Leahy, J. (1998), 'Miracle on Sixth Avenue: Information externalities and search', *Economic Journal* **108**(446), 60–74.

Carillo, P. E. & Rothbaum, J. L. (2016), 'Counterfactual spatial distributions', *Journal of Regional Science* **56**(5), 868–894.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. & Porter, S. R. (2018), 'The opportunity atlas: Mapping the childhood roots of social mobility', Working Paper 25147, National Bureau of Economic Research.

Conley, T. G. (1999), 'GMM estimation with cross sectional dependence', *Journal of Econometrics* **92**(1), 1–45.

Costa, D. & Kahn, M. (2000), 'Power couples: Changes in the locational choice of the college educated, 1940-1990', *Quarterly Journal of Economics* **115**(4), 1287–1315.

Couture, V., Gaubert, C., Handbury, J. & Hurst, E. (2018), 'Income growth and the distributional effects of urban spatial sorting', Mimeo.

Couture, V. & Handbury, J. (2017), 'Urban revival in America, 2000 to 2010', Working Paper 24084, National Bureau of Economic Research.

Ding, L., Hwang, J. & Divringi, E. (2016), 'Gentrification and residential mobility in Philadelphia', *Regional Science and Urban Economics* **61**(C), 38–51.

Duranton, G. (2007), 'Urban evolutions: The fast, the slow, and the still', *American Economic Review* **97**(1), 197–221.

Duranton, G. & Overman, H. G. (2005), 'Testing for localization using micro-geographic data', *Review of Economic Studies* **72**(4), 1077–1106.

Easterly, W., Freschi, L. & Pennings, S. (2018), '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.

Echenique, F. & Fryer, R. G. (2007), 'A measure of segregation based on social interactions', *The Quarterly Journal of Economics* **122**(2), 441–485.

Edlund, L., Machado, C. & Sviatschi, M. M. (2015), 'Bright minds, big rent: Gentrification and the rising returns to skill', Working Paper 21729, National Bureau of Economic Research.

Ellen, I. G., Horn, K. M. & Reed, D. (2017), 'Has falling crime invited gentrification?', Studies Paper CES-WP-17-27, US Census Bureau Center for Economic.

Fabrizio, C., Lalive, R., Sakalli, S. O. & Thoenig, M. (2018), 'Inference with arbitrary clustering', Mimeo, HEC University of Lausanne.

Freeman, L. (2005), 'Displacement or succession?', *Urban Affairs Review* **40**(4), 463–491.

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

Glaeser, E. L. & Kahn, M. E. (2001), 'Decentralized employment and the transformation of the American city', *Brookings-Wharton Papers on Urban Affairs* pp. 1–63.

Glaeser, E. L., Kim, H. & Luca, M. (2018), 'Measuring gentrification: Using Yelp data to quantify neighborhood change', Working Paper 24952, National Bureau of Economic Research.

Guerrieri, V., Hartley, D. & Hurst, E. (2013), 'Endogenous gentrification and housing price dynamics', *Journal of Public Economics* **100**(C), 45–60.

Hammel, D. J. & Wyly, E. K. (1996), 'A model for identifying gentrified areas with census data', *Urban Geography* **17**(3), 248–268.

Hidalgo, C. A. & Castañer, E. E. (2015), 'The amenity space and the evolution of neighborhoods', *ArXiv e-prints* .

Hwang, J. & Lin, J. (2016), 'What have we learned on the recent causes of gentrification?', *Cityscape: A Journal of Policy Development and Research* **18**(3), 9–26.

Kan, K., Kwong, S. K.-S. & Leung, C. K.-Y. (2004), 'The dynamics and volatility of commercial and residential property prices: Theory and evidence', *The Quarterly Journal of Economics* **44**(1), 95–123.

Lee, S. & Lin, J. (2018), 'Natural amenities, neighbourhood dynamics, and persistence in the spatial distribution of income', *Review of Economic Studies* **85**(1), 663–694.

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

Lester, T. W. & Hartley, D. A. (2014), 'The long term employment impacts of gentrification in the 1990s', *Regional Science and Urban Economics* **45**(C), 80–89.

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

McKinnish, T., Walsh, R. & Kirk White, T. (2010), 'Who gentrifies low-income neighborhoods?', *Journal of Urban Economics* **67**(2), 180–193.

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

Meltzer, R. & Ghorbani, P. (2017), 'Does gentrification increase employment opportunities in low-income neighborhoods?', *Regional Science and Urban Economics* **66**, 52 – 73.

Neumark, D., Wall, B. & Zhang, J. (2011), 'Do small businesses create more jobs? New evidence for the United States from the National Establishment Time Series', *The Review of Economics and Statistics* **93**(1), 16–29.

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

Pierce, J. R. & Schott, P. K. (2012), 'Concording U.S. harmonized system categories over time', *Journal of Official Statistics* **28**(1), 53–68.

Roback, J. (1982), 'Wages, rents, and the quality of life', *Journal of Political Economy* **90**(6), 1257–1278.

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

Rosenthal, S. S. & Ross, S. L. (2015), Chapter 16 - Change and persistence in the economic status of neighborhoods and cities, *in* J. V. H. Gilles Duranton & W. C. Strange, eds, 'Handbook of Regional and Urban Economics', Vol. 5 of *Handbook of Regional and Urban Economics*, Elsevier, pp. 1047 – 1120.

Schuetz, J. (2014), 'Do art galleries stimulate redevelopment?', *Journal of Urban Economics* **83**, 59 – 72.

Su, Y. (2018), 'The rising value of time and the origin of urban gentrification', Mimeo.

Sullivan, D. M. & Shaw, S. C. (2011), 'Retail gentrification and race: The case of Alberta street in Portland, Oregon', *Urban Affairs Review* .

Walls & Associates (n.d.), 'Technical documentation of the National Establishment Time Series database (NETS), year = 2014,'.

Zukin, S., Trujillo, V., Frase, P., Jackson, D., Recuber, T. & Walker, A. (2009), 'New retail capital and neighborhood change: Boutiques and gentrification in New York City', *City and Community* **8**(1), 47–64.

Appendix material

This set of appendices is structured as follows. Appendix **A** provides additional information on the data that we use. Appendix **B** explains in detail our methodology for concording large datasets over time. Appendix **C** details our procedure for clustering gentrifying blocks into 'neighborhoods'. Last, Appendix **D** provides additional tables and results.

Appendix A: Data

A.1. Additional information on data for the New York MSA. The area we consider as the New York 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 NHGIS data that we use for New York—and also for Philadelphia— are available at `https://www.nhgis.org`. All shape files we use are provided by the Census Tiger, and we use the 2010 versions. The algorithm described in Appendix B provides us with a concordance that allows us to associate the 2010 blocks with stable units. We then use those dissolved shape files to assign the NETS establishments to time-consistent blocks.

A.2. Additional data for Philadelphia. We 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 from the Southeastern Pennsylvania Transportation Authority (SEPTA) 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.

A.3. Additional information on the NETS data. Detailed information on the NETS data can be found in Walls & Associates (n.d.) and Neumark et al. (2011). Three important information for our analysis are the following.

First, as mentioned in the main text, the NETS data feature different qualities of geocoding. While much of the dataset is coded at the "block face" level, i.e., precisely, earlier years feature more observations where zip-code centroids are imputed. D&B underline that zip-codes may allow for more accurate positioning of businesses than census tracts or zip-code tabulation areas 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.

Second, as also explained in the may text, more recent years in the NETS data feature many more establishments. 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 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, this large increase should not be an issue.

Last, one may wonder how the NETS data compare with other establishment-level data for the U.S. 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).

Appendix B: Concording census blocks over time

We start with a simple example to explain our graph-theoretic approach to building concordances. Table 10 describes the structure of correspondence for a hypothetical nomenclature revised between years 1 and 2, and then 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 10, 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.

Observe that the correspondence in Table 10 has the same structure than the correspondence tables generally provided by statistical agencies (e.g., it is similar to that 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 concordance relationships are the edges. Being a graph, the correspondence in Table 10 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 9 displays the graph associated with Table 10. Each node (e.g., a1 or d3) corresponds

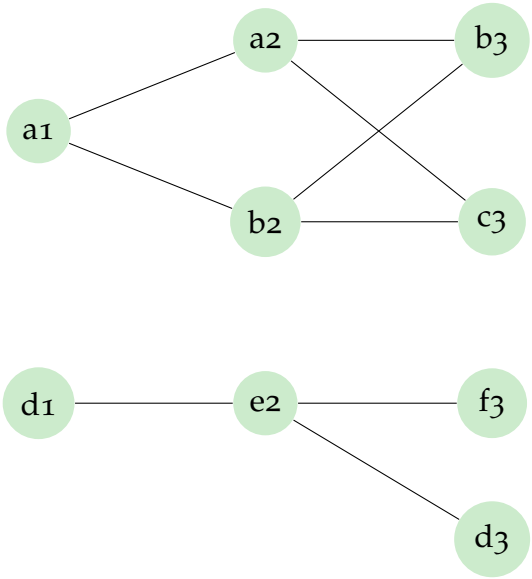
Table 10: Sample correspondence table.

		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 of a correspondence table 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).

to a unique code-year combination. As can be seen, optimally concording the codes of Table 10 requires finding the smallest groups of codes that are all connected and thus define components that are invariant and comparable over time. Figure 9 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 smallest time-invariant synthetic groups 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).

Figure 9: Example of connected components.



Any concordance problem based on crosswalks provided by statistical agencies can be viewed as in Table 10 and Figure 9. 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:

Algorithm 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 'smallest synthetic groups') 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. This method is simple, extremely efficient, universally applicable, produces minimum concordances, and can be readily implemented using standard 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).[31] We use `Stata` to prepare the intermediate data, and `networkComponents.m`, an open source `Matlab` code developed by Daniel Larremore, to find the connected components of the graph.

Appendix C: Clustering blocks into neighborhoods

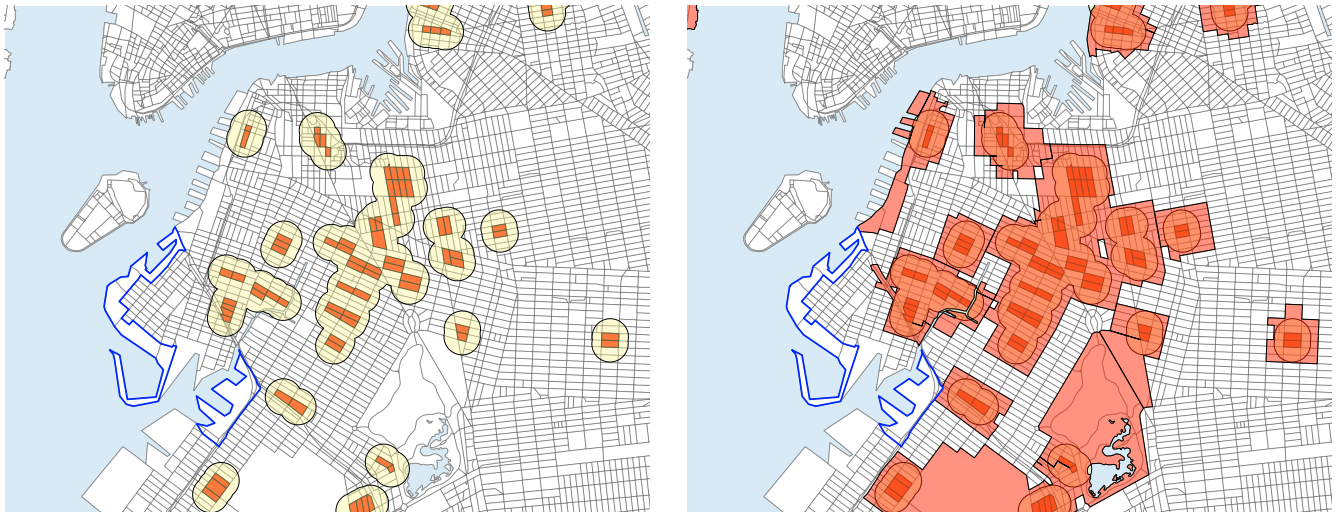
We move from blocks to neighborhoods by identifying areas with clusters of gentrifying blocks. The idea is to identify larger zones where rapid socio-economic change takes place, and to link those zones more closely with areas that have been written about extensively in the media. Our cluster procedure draws on tools from spatial point-pattern analysis. We first compute, for each gentrifying block i, the number of *all other* blocks and the number of *non-gentrifying* blocks in a radius of 250 meters around block i. Assume that there are n_i gentrifying blocks and m_i non-gentrifying blocks in that radius. If there are N gentrifying blocks and M non-

[31] There are, e.g., several papers that concord product nomenclatures over time (see Pierce & Schott 2012 for U.S. product categories; Martin & 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 & Schott (2012). They are hence not portable. Tests that we ran also suggest that they are much slower than our approach for large datasets.

gentrifying blocks in total in the New York metropolitan area between two census years, then the probability that there are more than n_i gentrifying blocks among the $n_i + m_i$ total blocks around i can be computed from the cumulative distribution function of a hypergeometric distribution. Assume that this value is 0.01 for block i. This means that there is only a 1% chance of observing more than n_i gentrifying blocks around block i, conditional on having $n_i + m_i$ blocks in total around block i and conditional on the overall share $N/(N+M)$ of gentrifying blocks in New York. In words, we are very unlikely to observe that many gentrifying blocks around i, i.e., there is clustering of gentrifying blocks around the *focal block* i.[32]

Figure 10: Focal block buffers and gentrifying neighborhoods, 1990–2000.

(a) Focal block buffers. (b) Gentrifying neighborhoods.



Keeping all *focal blocks*—defined as gentrifying blocks with a p-value below 0.01—we then draw 250 meter buffers around those focal blocks and take their unions to produce disjoint areas that we call 'gentrifying neighborhoods'. This procedure is illustrated in Figure 10. We finally add all blocks that intersect with those buffers to the neighborhood. Panel (b) of Figure 10 depicts the neighborhoods we identify as gentrifying in northern Brooklyn between 1990 and 2000.

Appendix D: Additional tables and figures

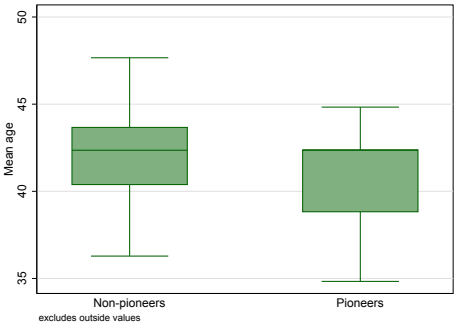
This appendix provides additional figures and estimation results. Figure 11 summarizes characteristics of workers in pioneer industries for the whole U.S. Tables 11 and 12 provide estima-

[32]This counterfactual is similar in spirit to the one used in, e.g., Duranton & Overman (2005). It corresponds to a random reshuffling of block types—gentrifying and non-gentrifying blocks—across all blocks in New York. Contrary to Duranton & Overman (2005), we do not control for the correlation between successive draws when computing the p-values, i.e., we assume that draws for each block i are independent. This should matter little in practice as the number of gentrifying blocks is just about 5% of all blocks.

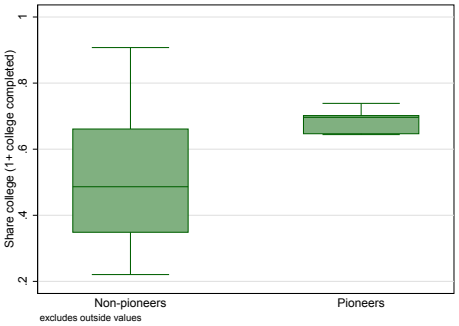
tion results for the predictions of gentrifying blocks in New York using 500 meters rings and a block-contiguity matrix to construct the explanatory variables, respectively. Table 13 contains regression results on the performance of pioneers and non-pioneers in gentrifying areas.

Figure 11: Worker characteristics in pioneer and non-pioneer industries in the U.S.

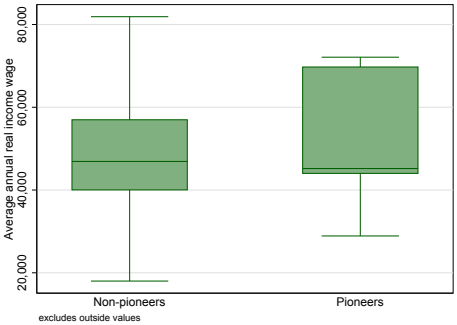
(a) Pioneers are younger.



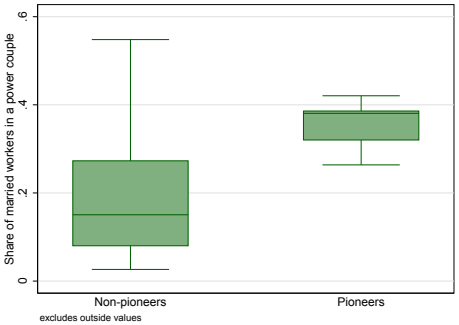
(b) Pioneers are more educated.



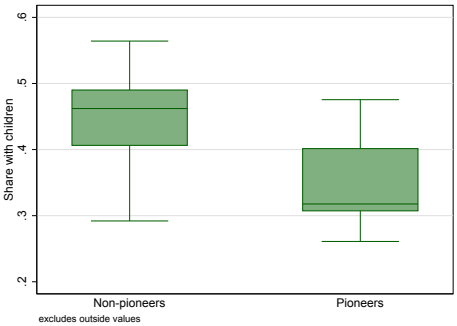
(c) Pioneers are not much wealthier.



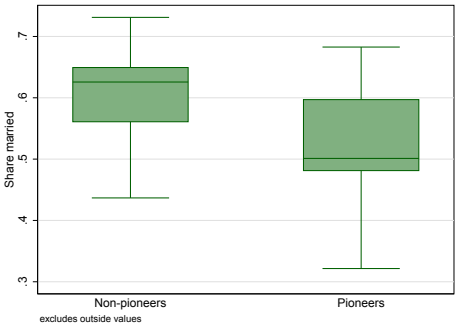
(d) Pioneers are in 'power couples'.



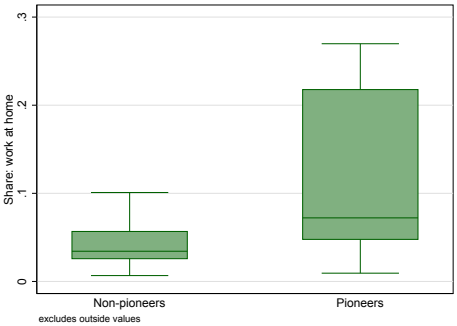
(e) Pioneers have fewer children.



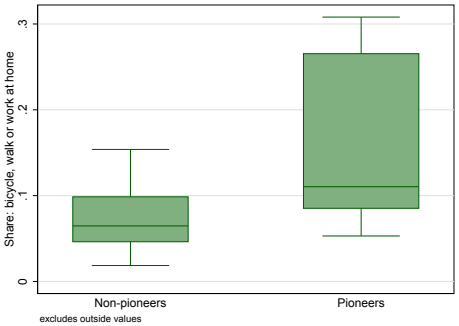
(f) Pioneers are more often single.



(g) Pioneers work more at home.



(h) Pioneers commute shorter distances.



Notes: Our computations using IPUMS data for the years 2000 and 2010.

Table 11: Determinants of gentrification in New York and Philadelphia, 2000–2010 (500 meters radius).

Dependent variable	Dummy equal to one blocks that gentrify during 2000–2010, $\mathbb{1}_b^{gentri}$.					
	New York				Philadelphia	
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to pioneers	0.053^a	0.055^a	0.054^a	0.051^a	0.053^a	0.090^a
	(0.008)	(0.009)	(0.013)	(0.017)	(0.011)	(0.021)
Number of establishments (log)	-0.020^a	-0.014	-0.028^a	-0.018	-0.005	-0.043^a
	(0.007)	(0.010)	(0.010)	(0.014)	(0.010)	(0.016)
Per capita income (log)	0.029	0.039	0.027	0.042	0.090^b	0.076^b
	(0.021)	(0.029)	(0.021)	(0.029)	(0.040)	(0.038)
Population (log)	-0.005	-0.025^b	-0.002	-0.020^b	-0.036^a	-0.029^b
	(0.007)	(0.010)	(0.006)	(0.010)	(0.013)	(0.013)
Share college educated	0.446^a	0.506^b	0.441^a	0.519^b	0.213	0.153
	(0.165)	(0.221)	(0.170)	(0.227)	(0.151)	(0.153)
Share African-American	-0.038	-0.101^c	-0.049	-0.109^c	-0.187^b	-0.203^a
	(0.042)	(0.061)	(0.041)	(0.061)	(0.075)	(0.073)
Share Asian	-0.324^b	-0.504^a	-0.315^b	-0.499^a	0.037	-0.005
	(0.135)	(0.151)	(0.138)	(0.154)	(0.479)	(0.483)
Share other minority residents	-0.167^c	-0.139	-0.164	-0.148	0.085	0.065
	(0.101)	(0.132)	(0.102)	(0.134)	(0.185)	(0.184)
Rent (log)	-0.020	-0.011	-0.021	-0.014	-0.011	-0.013
	(0.014)	(0.018)	(0.014)	(0.019)	(0.028)	(0.028)
Median age of buildings (log)	0.001^b	0.002^a	0.001^a	0.002^a	0.004^a	0.004^a
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Dist. to closest gentrifying block 1990–2000 (log)	-0.010^b	-0.007	-0.011^b	-0.007	-0.025^b	-0.021^b
	(0.004)	(0.005)	(0.004)	(0.006)	(0.010)	(0.010)
Less than 200m from waterfront	0.013	0.026	0.012	0.024	0.002	-0.003
	(0.013)	(0.023)	(0.013)	(0.023)	(0.021)	(0.022)
Distance to subway (log)	-0.009^a	-0.002	-0.010^a	-0.003	-0.001	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
Distance to closest park (log)	-0.005^b	-0.007^a	-0.004^c	-0.007^b	0.002	0.003
	(0.002)	(0.003)	(0.002)	(0.003)	(0.005)	(0.005)
# of main landmarks (log)	-0.000	0.012	0.000	0.013	0.001	0.000
	(0.007)	(0.009)	(0.007)	(0.009)	(0.012)	(0.012)
Robbery		0.033^a		0.034^a		
		(0.006)		(0.006)		
Burglary		-0.016^a		-0.016^a		
		(0.006)		(0.006)		
Murder		-0.203		-0.210		
		(0.150)		(0.147)		
Rape		-0.109		-0.117		
		(0.111)		(0.112)		
# of observations	34,164	19,868	34,164	19,868	18,144	18,144
Sample	New York	NYC	New York	NYC	Philadelphia	Philadelphia
Specification	OLS-HAC	OLS-HAC	IV	IV	OLS-HAC	IV

Notes: The sample is composed of blocks which income per capita is below the median in the city in 2000. The measure of exposure to positive pioneers is given by (4). All explanatory variables are measured in 2000 and are computed using 250 meter rings around each block (except the distance to subway, to parks and to closest gentrifying block variables, as well as the waterfront dummy). Standard errors corrected for cross-sectional spatial dependence within a 500-meter radius (from HAC estimation) in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

Table 12: Determinants of gentrification in New York and Philadelphia, 2000–2010 (contiguous blocks).

Dependent variable	Dummy equal to one for blocks that gentrify during 2000–2010, $\mathbb{1}_b^{gentri}$.					
	New York				Philadelphia	
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to pioneers	0.054^a	0.057^a	0.056^a	0.059^a	0.057^a	0.048^b
	(0.008)	(0.010)	(0.012)	(0.016)	(0.013)	(0.022)
Number of establishments (log)	-0.009^b	-0.007	-0.012^a	-0.009	0.000	0.000
	(0.004)	(0.005)	(0.005)	(0.006)	(0.005)	(0.007)
Per capita income (log)	0.001	0.038	0.000	0.037	-0.033	-0.034
	(0.018)	(0.024)	(0.018)	(0.024)	(0.026)	(0.026)
Population (log)	-0.008	-0.012^b	-0.007	-0.012^b	-0.037^a	-0.037^a
	(0.005)	(0.006)	(0.005)	(0.006)	(0.010)	(0.010)
Share college educated	0.318^a	0.218^b	0.314^a	0.210^b	0.356^a	0.376^a
	(0.078)	(0.100)	(0.080)	(0.101)	(0.131)	(0.133)
Share African-American	-0.010	-0.036^c	-0.011	-0.037^c	-0.099^a	-0.100^a
	(0.014)	(0.021)	(0.014)	(0.020)	(0.023)	(0.023)
Share Asian	-0.096^b	-0.152^a	-0.096^b	-0.152^a	-0.062	-0.059
	(0.040)	(0.044)	(0.041)	(0.044)	(0.146)	(0.146)
Share other minority residents	-0.026	-0.043	-0.027	-0.044	-0.104^b	-0.105^b
	(0.040)	(0.048)	(0.040)	(0.048)	(0.048)	(0.048)
Rent (log)	-0.026^c	-0.028	-0.027^c	-0.028	-0.005	-0.005
	(0.015)	(0.019)	(0.015)	(0.020)	(0.025)	(0.025)
Median age of buildings (log)	0.001^a	0.002^a	0.001^a	0.002^a	0.003^a	0.002^a
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Dist. to closest gentrifying block 1990–2000 (log)	-0.012^a	-0.009^c	-0.012^a	-0.009^c	-0.031^a	-0.032^a
	(0.004)	(0.006)	(0.004)	(0.006)	(0.010)	(0.010)
Less than 200m from waterfront	0.014	0.031	0.014	0.031	0.010	0.010
	(0.013)	(0.023)	(0.013)	(0.023)	(0.019)	(0.019)
Distance to subway (log)	-0.009^a	-0.004	-0.010^a	-0.004	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
Distance to closest park (log)	-0.005^b	-0.008^a	-0.004^c	-0.008^a	0.002	0.002
	(0.002)	(0.003)	(0.002)	(0.003)	(0.005)	(0.005)
# of main landmarks (log)	-0.008	-0.001	-0.007	-0.001	-0.004	-0.003
	(0.006)	(0.010)	(0.006)	(0.010)	(0.011)	(0.011)
Robbery		0.033^a		0.033^a		
		(0.006)		(0.006)		
Burglary		-0.019^a		-0.019^a		
		(0.006)		(0.006)		
Murder		-0.224		-0.223		
		(0.150)		(0.149)		
Rape		-0.100		-0.101		
		(0.108)		(0.109)		
# of observations	34,164	19,848	34,164	19,848	18,144	18,144
Sample	New York	NYC	New York	NYC	Philadelphia	Philadelphia
Specification	OLS-HAC	OLS-HAC	IV	IV	OLS-HAC	IV

Notes: The sample is composed of blocks which income per capita is below the median in the city in 2000. The measure of exposure to positive pioneers is given by (4). All explanatory variables are measured in 2000 and are computed using 250 meter rings around each block (except the distance to subway, to parks and to closest gentrifying block variables, as well as the waterfront dummy). Standard errors corrected for cross-sectional spatial dependence within a 500-meter radius (from HAC estimation) in parentheses. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.

Table 13: Evolution of the performance of establishments, 1990–2000.

Dependent variable	Employment	Credit score	Survival	Employment	Credit score	Survival
	(1)	(2)	(3)	(4)	(5)	(6)
Employment in 1990 (log)	-0.077^a	0.032^a	0.011^a	-0.077^a	0.032^a	0.011^a
	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Gentrification	0.012^a	-0.007	-0.019^a	0.012^a	-0.007	-0.019^a
	(0.004)	(0.009)	(0.003)	(0.004)	(0.009)	(0.003)
Gentrification × Pioneer industry	0.029	-0.020	0.008	0.031	-0.022	0.013
	(0.018)	(0.040)	(0.010)	(0.019)	(0.041)	(0.011)
Gentrification × Cons. amenity				-0.028	0.030	-0.069^c
				(0.055)	(0.151)	(0.039)
Fixed effects	6-digit NAICS industries					
Observations	173,475	34,809	375,988	173,475	34,809	375,988
R^2	0.041	0.051	0.059	0.041	0.051	0.059

Notes: The sample is composed of all establishments in 1990 in New York. All dependent variables are expressed as the change in the establishment's performance (level of employment and 'Duns & Bradstreet's Paydex score') between 1990 and 2000, except 'Survival' which values one if the establishments is still active in 2000, and zero otherwise. 'Gentrification' is a dummy variable equal to one if the block gentrifies between 1990 and 2000. 'Pioneer industry' is a dummy variable equal to one if the establishment is active in a pioneer industry. 'Cons. amenity' is a dummy variable equal to one if the establishment both (i) belongs to a pioneer sector and (ii) is a consumption amenity, as defined in Couture & Handbury (2017). The latter industries are 'Promoters of Performing Arts, Sports, and Similar Events without facilities', 'Theater Companies and Dinner Theaters', and 'Museums'. All regressions include 6-digit NAICS industries fixed effects. Huber-White robust standard errors in parenthesis. ^{a} = significant at 1%, ^{b} = significant at 5%, ^{c} = significant at 10%.