

# Gentrification and Pioneer Businesses:

## Supplemental Online Appendix

Kristian Behrens   Brahim Boualam   Julien Martin   Florian Mayneris

### O.1. Harmonized blocks

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

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

---

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

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

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

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

---

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

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

## O.2. Descriptive statistics.

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

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

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

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

## O.3. Changes in income and education.

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

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

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

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

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

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

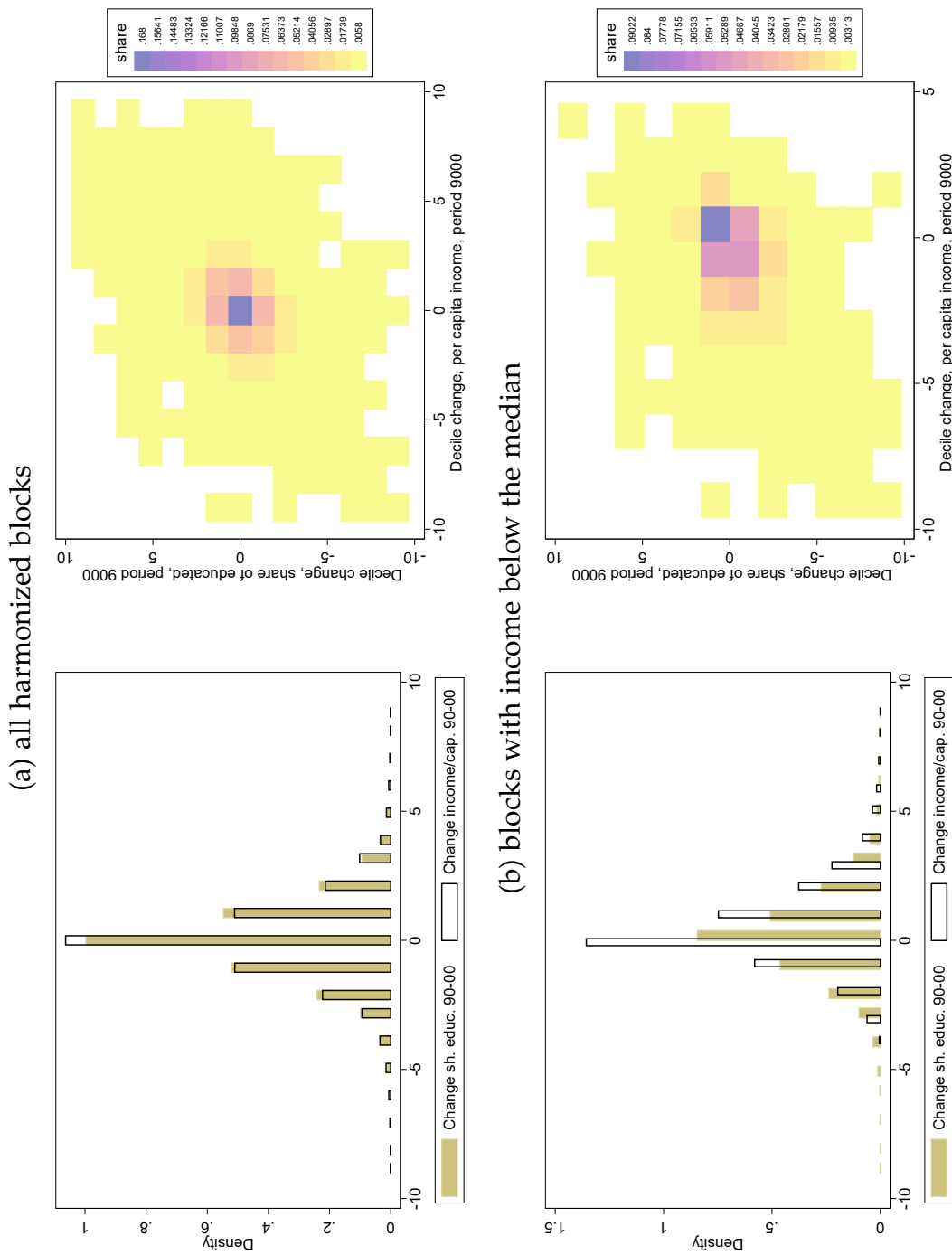


Table O.2: Contribution of between-tract variations to overall block-level variations in income and education.

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

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

and their connection with the presence of pioneer businesses.

## O.4. Bibliometric analysis.

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

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

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

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

## O.5. Additional information on NETS data.

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

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

---

<sup>29</sup>See [Walls and Associates \(2014\)](#) and [Neumark et al. \(2011\)](#) for more information on NETS data.

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

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

(a) Accuracy of geocoding						
Geocoding type	Share of establishments			Share of employment		
	1990	2000	2010	1990	2000	2010
Block face	73.4%	85.9%	96.6%	77.4%	87.7%	94.4%
ZIP-code	25.5%	12.8%	2.1%	20.3%	9.5%	2.9%
Others	1.1%	1.3%	1.3%	2.3%	2.8%	2.7%

(b) Establishment size distribution			
# of employees	1990	2000	2010
1	108,735	200,569	376,629
2 to 5	329,214	439,527	671,104
6 to 10	96,368	103,409	97,080
11 to 50	95,810	105,606	99,927
50+	25,869	27,524	26,981
Total	655,996	876,635	1,271,721

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

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

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

## O.6. Geographic concordance algorithm.

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

Table O.5: Sample correspondence table.

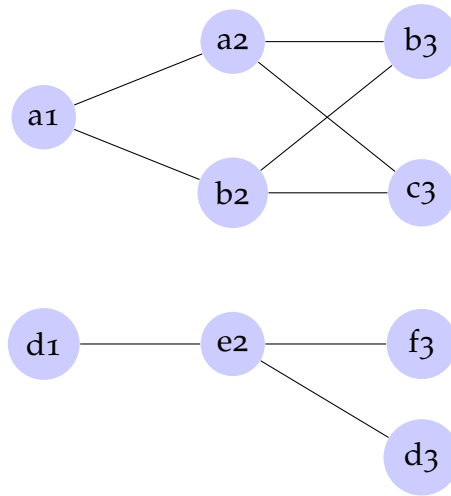
Years	Obs	Old			New		
		Code	Partial flag	Year	Code	Partial flag	Year
1-2	[1]	a	<i>p</i>	1	a		2
1-2	[2]	a	<i>p</i>	1	b		2
1-2	[3]	d		1	e		2
2-3	[4]	b	<i>p</i>	2	b	<i>p</i>	3
2-3	[5]	b	<i>p</i>	2	c	<i>p</i>	3
2-3	[6]	a	<i>p</i>	2	b	<i>p</i>	3
2-3	[7]	a	<i>p</i>	2	c	<i>p</i>	3
2-3	[8]	e	<i>p</i>	2	f		3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

*Notes:* Example with three years. Statistical agencies would provide one table for the passage from year 1 to 2 (top panel), and a separate table for the passage from year 2 to 3 (bottom panel).

'Partial flag' identifies modifications that are not 1 to 1.

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

Figure O.2: Example of connected components.



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

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

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

---

**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 SSG) corresponds simply to finding the connected components of the graph spanned by the code-year nodes and the revision edges. Once viewed in these terms, it becomes a relatively standard problem that can be solved efficiently using the tools of graph theory to find the connected components and to build synthetic identifiers for related codes.<sup>31</sup> This method is simple, extremely efficient, universally applicable, produces minimum concordances, and can be readily implemented using standard

---

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

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

## O.7. Descriptives of harmonized blocks

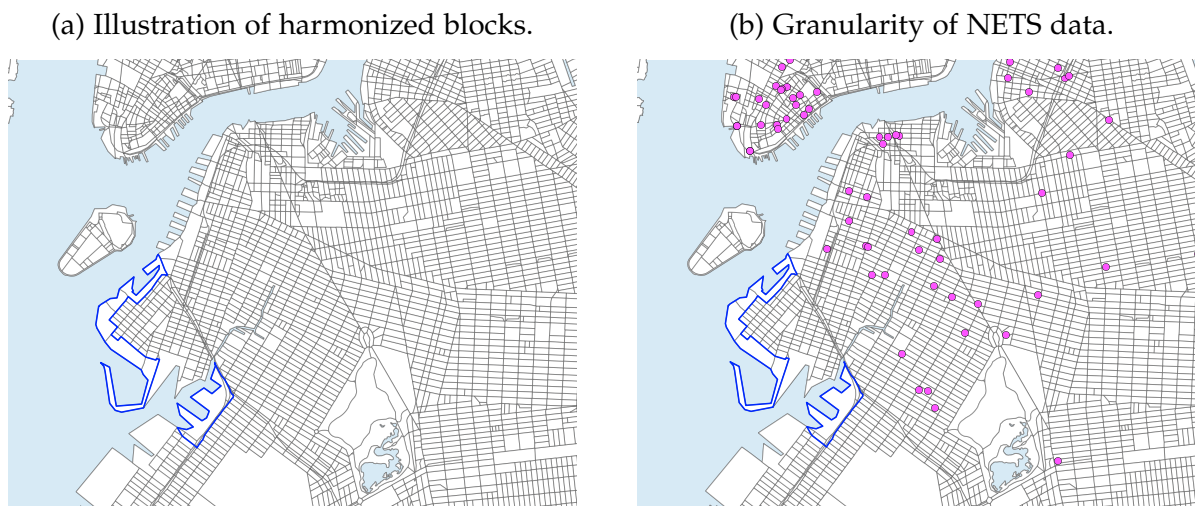
As is well known the smallest synthetic groups (SSGs) are, by definition, more aggregated than the original units from which they are constructed. This naturally raises the question of how much geographic information we lose when harmonizing census blocks over time. Our algorithm identifies 152,529 time-consistent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

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

---

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

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



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

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

---

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

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

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

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

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

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

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

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

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

## O.9. Additional tables and figures

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

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

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

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

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

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
$\Delta$ # pioneer estab.	1.406 <sup>a</sup>	0.959 <sup>a</sup>	1.852 <sup>a</sup>	1.484 <sup>a</sup>	0.378 <sup>a</sup>	0.241 <sup>a</sup>
	(0.351)	(0.254)	(0.420)	(0.324)	(0.127)	(0.090)
$\Delta$ Ln (1+ # non pioneer estab)	0.011	0.000	0.018	0.007	0.012 <sup>a</sup>	0.001
	(0.010)	(0.014)	(0.013)	(0.015)	(0.004)	(0.006)
Ln per cap. income	0.045 <sup>b</sup>	0.057 <sup>b</sup>	-0.090 <sup>a</sup>	-0.065 <sup>c</sup>	0.059 <sup>a</sup>	0.070 <sup>a</sup>
	(0.020)	(0.028)	(0.029)	(0.039)	(0.007)	(0.010)
Share college edu. resid.	0.598 <sup>a</sup>	0.535 <sup>b</sup>	1.111 <sup>a</sup>	0.494 <sup>b</sup>	0.041	-0.084
	(0.165)	(0.225)	(0.181)	(0.234)	(0.070)	(0.089)
Ln rent	-0.022	-0.008	-0.020	-0.003	-0.013 <sup>b</sup>	-0.007
	(0.014)	(0.018)	(0.017)	(0.022)	(0.005)	(0.007)
Median age of buildings	0.001 <sup>a</sup>	0.002 <sup>a</sup>	0.001	0.001	0.000 <sup>b</sup>	0.001 <sup>a</sup>
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Share black resid.	-0.061	-0.033	-0.115 <sup>b</sup>	0.005	0.015	0.041 <sup>c</sup>
	(0.041)	(0.061)	(0.051)	(0.075)	(0.016)	(0.023)
Share asian resid.	-0.519 <sup>a</sup>	-0.581 <sup>a</sup>	-0.702 <sup>a</sup>	-0.708 <sup>a</sup>	-0.076	-0.148 <sup>a</sup>
	(0.129)	(0.143)	(0.139)	(0.160)	(0.050)	(0.055)
Share other resid.	-0.167 <sup>c</sup>	-0.176	-0.706 <sup>a</sup>	-0.246	0.019	0.089 <sup>c</sup>
	(0.100)	(0.130)	(0.143)	(0.156)	(0.041)	(0.049)
Ln pop.	-0.002	-0.036 <sup>a</sup>	0.009	-0.053 <sup>a</sup>	0.005 <sup>b</sup>	-0.017 <sup>a</sup>
	(0.006)	(0.013)	(0.008)	(0.016)	(0.002)	(0.005)
Less than 200m from waterfront	0.015	0.022	0.015	0.039 <sup>c</sup>	0.005	0.005
	(0.013)	(0.024)	(0.014)	(0.021)	(0.005)	(0.007)
Ln (1+# train lines)	0.020 <sup>a</sup>	0.005	0.024 <sup>a</sup>	0.005	0.009 <sup>a</sup>	0.002
	(0.006)	(0.007)	(0.007)	(0.008)	(0.002)	(0.002)
Ln (1+# bus lines)	-0.003	-0.012	-0.006	-0.001	-0.001	-0.003
	(0.005)	(0.011)	(0.005)	(0.010)	(0.002)	(0.003)
Distance to closest park (log)	-0.005 <sup>b</sup>	-0.008 <sup>a</sup>	-0.002	-0.004	-0.001	-0.002
	(0.002)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)
Ln # of main landmarks	-0.001	0.016 <sup>c</sup>	0.005	0.016	0.004	0.006
	(0.007)	(0.009)	(0.008)	(0.010)	(0.003)	(0.003)
Socio-economic changes in the neighborhood 1990–2000	-0.014 <sup>a</sup>	-0.011 <sup>b</sup>	0.098 <sup>b</sup>	0.051	-0.709 <sup>a</sup>	-0.515 <sup>a</sup>
	(0.004)	(0.006)	(0.043)	(0.053)	(0.124)	(0.167)
# murder per cap.		-0.370 <sup>b</sup>		-0.494 <sup>b</sup>		-0.085
		(0.153)		(0.201)		(0.064)
# burglary per cap.		-0.020 <sup>a</sup>		-0.029 <sup>a</sup>		-0.013 <sup>a</sup>
		(0.006)		(0.008)		(0.003)
# robbery per cap.		0.034 <sup>a</sup>		0.052 <sup>a</sup>		0.014 <sup>a</sup>
		(0.006)		(0.007)		(0.002)
# rape per cap.		-0.103		-0.207		-0.050
		(0.109)		(0.128)		(0.041)
Ln (1+# rent control buildings)		0.014 <sup>a</sup>		0.013 <sup>a</sup>		0.008 <sup>a</sup>
		(0.005)		(0.005)		(0.002)
Share vacant land		0.044		-0.040		-0.029
		(0.162)		(0.226)		(0.048)
Presence of limited height districts		0.072		-0.081		0.011
		(0.264)		(0.098)		(0.042)
Presence of historical districts		0.014		0.028 <sup>c</sup>		0.010 <sup>c</sup>
		(0.015)		(0.017)		(0.006)
# of observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.044	0.074	0.039	0.069	0.043	0.074
Sample	New York	NYC	New York	NYC	New York	NYC

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

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

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
$\Delta$ # pioneer estab.	0.007 <sup>a</sup> (0.002)	0.014 <sup>b</sup> (0.006)	0.009 <sup>a</sup> (0.002)	0.022 <sup>b</sup> (0.009)	0.001 <sup>c</sup> (0.001)	0.004 <sup>b</sup> (0.002)
$\Delta$ # non-pioneer estab.	0.002 (0.018)	-0.007 (0.024)	0.035 (0.024)	0.018 (0.036)	0.004 (0.008)	0.001 (0.010)
Ln per cap. income	0.103 <sup>a</sup> (0.034)	0.081 <sup>b</sup> (0.034)	-0.087 <sup>b</sup> (0.044)	-0.134 <sup>a</sup> (0.050)	0.045 <sup>a</sup> (0.013)	0.034 <sup>a</sup> (0.013)
Share college edu. resid.	0.249 <sup>c</sup> (0.137)	0.189 (0.140)	0.278 <sup>c</sup> (0.165)	0.157 (0.177)	-0.131 <sup>b</sup> (0.057)	-0.147 <sup>b</sup> (0.058)
Ln rent	-0.015 (0.026)	-0.012 (0.026)	-0.027 (0.026)	-0.023 (0.026)	-0.017 <sup>b</sup> (0.008)	-0.016 <sup>b</sup> (0.008)
Median age of buildings	0.004 <sup>a</sup> (0.001)	0.003 <sup>a</sup> (0.001)	0.000 (0.001)	-0.001 (0.001)	0.001 <sup>a</sup> (0.000)	0.001 <sup>a</sup> (0.000)
Share black resid.	-0.180 <sup>a</sup> (0.063)	-0.157 <sup>b</sup> (0.066)	-0.197 <sup>b</sup> (0.083)	-0.148 (0.092)	-0.043 (0.029)	-0.037 (0.030)
Share asian resid.	-0.048 (0.412)	-0.398 (0.389)	0.316 (0.527)	-0.391 (0.513)	0.030 (0.093)	-0.104 (0.113)
Share other resid.	0.056 (0.150)	0.006 (0.150)	-0.578 <sup>c</sup> (0.312)	-0.650 <sup>b</sup> (0.314)	-0.015 (0.075)	-0.032 (0.076)
Ln pop.	-0.033 <sup>a</sup> (0.011)	-0.047 <sup>a</sup> (0.014)	-0.056 <sup>a</sup> (0.014)	-0.086 <sup>a</sup> (0.022)	-0.013 <sup>a</sup> (0.004)	-0.018 <sup>a</sup> (0.005)
Less than 200m from waterfront	-0.001 (0.019)	0.003 (0.020)	-0.019 (0.016)	-0.012 (0.017)	0.000 (0.005)	0.002 (0.005)
Ln Distance to subway	-0.002 (0.005)	-0.002 (0.005)	-0.003 (0.005)	-0.002 (0.005)	-0.002 (0.002)	-0.002 (0.002)
Ln Distance to closest park	0.002 (0.005)	0.001 (0.005)	0.005 (0.004)	0.003 (0.004)	0.001 (0.001)	0.000 (0.001)
Ln (1+# of main landmarks)	0.002 (0.012)	-0.001 (0.011)	0.001 (0.014)	-0.005 (0.014)	0.001 (0.004)	-0.001 (0.004)
Socio-economic changes in the neighborhood 1990–2000	-0.024 <sup>a</sup> (0.009)	-0.017 <sup>c</sup> (0.009)	-0.010 (0.056)	-0.046 (0.062)	-0.585 <sup>a</sup> (0.106)	-0.598 <sup>a</sup> (0.109)
# of observations	18,144	18,144	18,009	18,009	18,014	18,014
R-squared	0.079	n.a.	0.040	n.a.	0.053	n.a.
Specification	LPM	IV	OLS	IV	OLS	IV

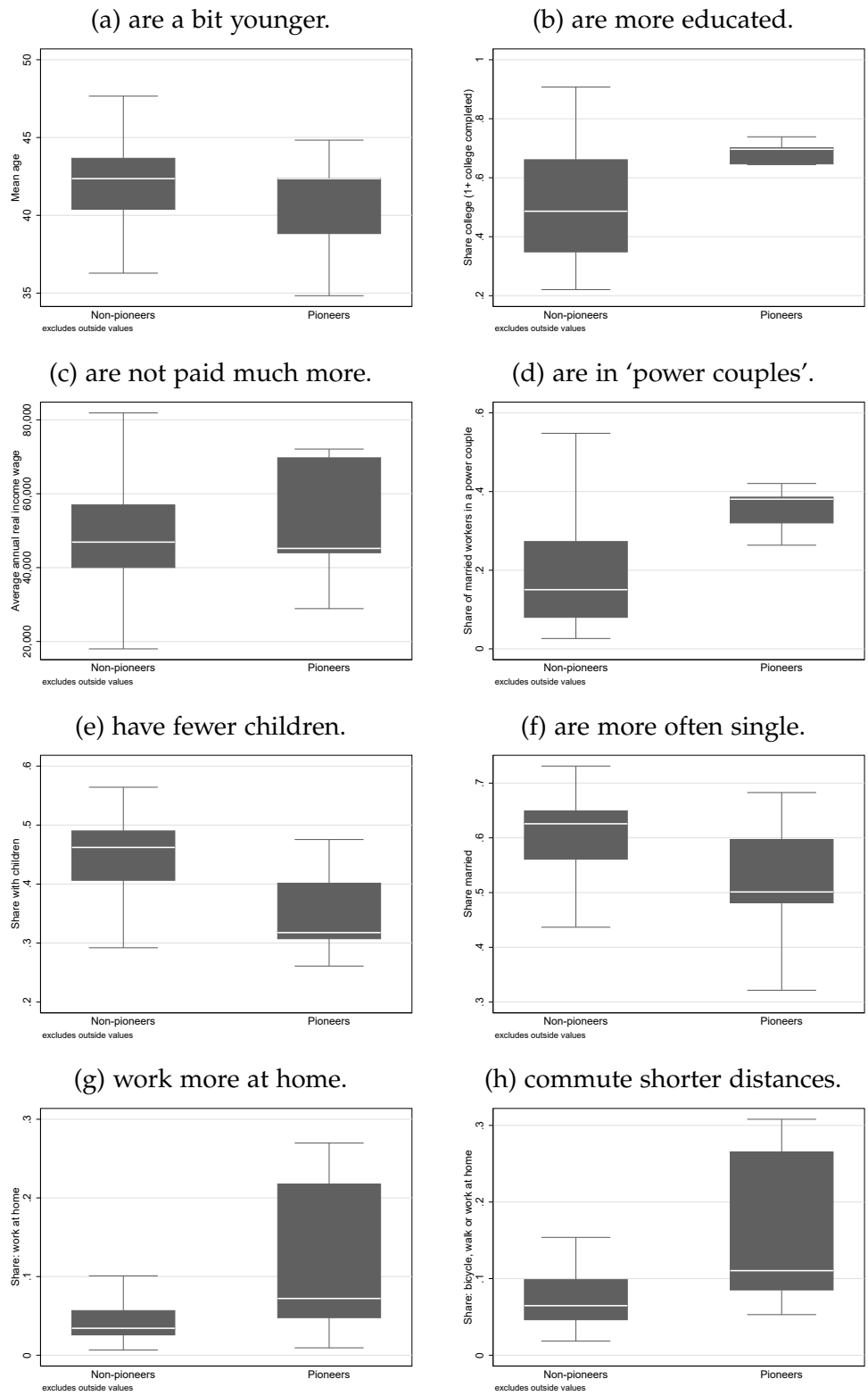
Notes: The sample is composed of blocks with per capita income below the median in the city in 2000, and with at least eight residents. The measure of exposure to pioneers is given by equation (3) in the main text. All explanatory variables are measured in 2000 and are computed using 500 meters rings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

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

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
(a) Interaction with prime locations						
$\Delta$ # pioneer estab.	1.106 <sup>a</sup> (0.342)	0.115 (0.231)	1.694 <sup>a</sup> (0.488)	0.709 <sup>b</sup> (0.310)	0.351 <sup>b</sup> (0.158)	0.031 (0.113)
— × log dist. prime locations	0.325 (0.263)	0.598 <sup>b</sup> (0.297)	0.127 (0.307)	0.610 <sup>c</sup> (0.350)	0.019 (0.123)	0.074 (0.134)
log dist. prime locations	-9.287 (6.109)	-109.315 <sup>a</sup> (17.519)	-9.233 (7.052)	-99.181 <sup>a</sup> (20.318)	-1.858 (2.549)	-38.317 <sup>a</sup> (6.827)
Controls	✓	✓	✓	✓	✓	✓
Observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.045	0.087	0.039	0.076	0.043	0.086
Sample	New York	NYC	New York	NYC	New York	NYC
(b) Interaction with waterfront						
$\Delta$ # pioneer estab.	1.387 <sup>a</sup> (0.323)	0.968 <sup>a</sup> (0.229)	1.834 <sup>a</sup> (0.374)	1.506 <sup>a</sup> (0.281)	0.372 <sup>a</sup> (0.121)	0.241 <sup>a</sup> (0.087)
— × log dist. waterfront	-0.354 <sup>b</sup> (0.142)	-0.490 <sup>a</sup> (0.186)	-0.557 <sup>a</sup> (0.195)	-0.645 <sup>b</sup> (0.306)	-0.094 <sup>c</sup> (0.054)	-0.097 (0.064)
log dist. waterfront	0.011 (0.798)	0.277 (1.192)	0.991 (0.822)	-0.102 (1.131)	-0.060 (0.242)	-0.088 (0.321)
Controls	✓	✓	✓	✓	✓	✓
Observations	34,022	19,947	33,718	19,766	33,725	19,772
R-squared	0.045	0.076	0.040	0.071	0.044	0.075
Sample	New York	NYC	New York	NYC	New York	NYC

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

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



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

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

	(1)	(2)	(3)	(4)	(5)	(6)
	Gentrification indicator		Change, ln per capita income		Change, share of educated	
$\Delta$ # pioneer estab.	0.002 <sup>a</sup> (0.001)	0.002 <sup>a</sup> (0.001)	0.003 <sup>a</sup> (0.001)	0.003 <sup>a</sup> (0.001)	0.001 <sup>a</sup> (0.000)	0.001 <sup>a</sup> (0.000)
$\Delta$ # other creative plants	-0.002 <sup>c</sup> (0.001)	-0.002 (0.001)	-0.004 <sup>a</sup> (0.001)	-0.003 <sup>b</sup> (0.001)	-0.001 <sup>a</sup> (0.000)	-0.001 <sup>b</sup> (0.001)
$\Delta$ # plants with similar charact.	-0.003 (0.002)	-0.004 <sup>c</sup> (0.002)	-0.003 (0.002)	-0.003 (0.002)	-0.001 (0.001)	-0.001 (0.001)
Controls	✓	✓	✓	✓	✓	✓
Observations	34,164	20,005	33,856	19,822	33,863	19,828
R-squared	0.045	0.087	0.039	0.076	0.043	0.086
Sample	New York	NYC	New York	NYC	New York	NYC

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

## O.10. Geography of gentrification

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

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

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

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

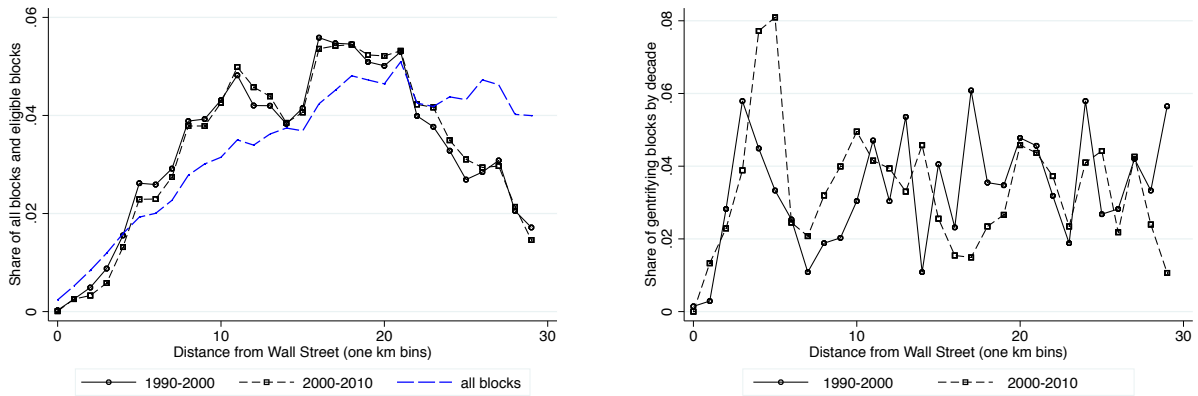
---

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

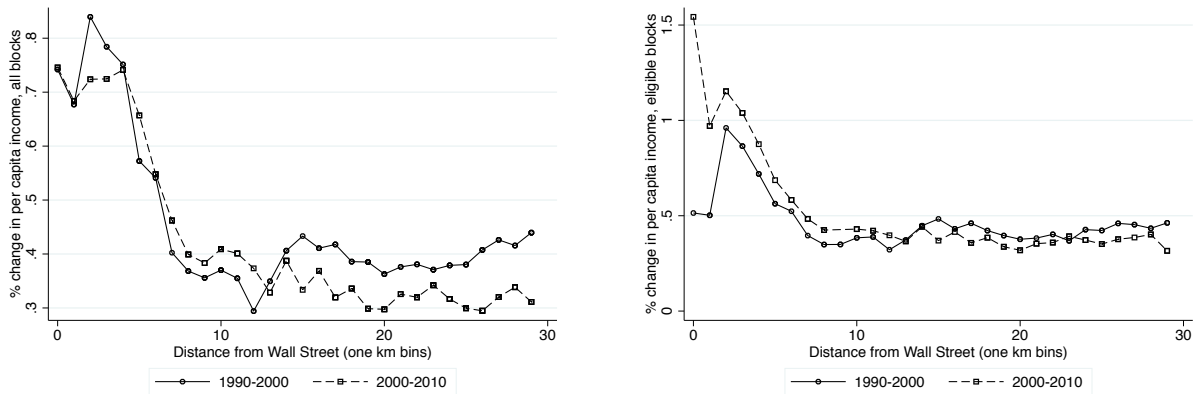
<sup>35</sup>Gentrification is not exclusively a central-city phenomenon: there are many gentrifying areas that are not close to the city center. A substantial amount of gentrification seems to partly follow the distribution of new housing, which is either located in the central city (due to renewal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

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

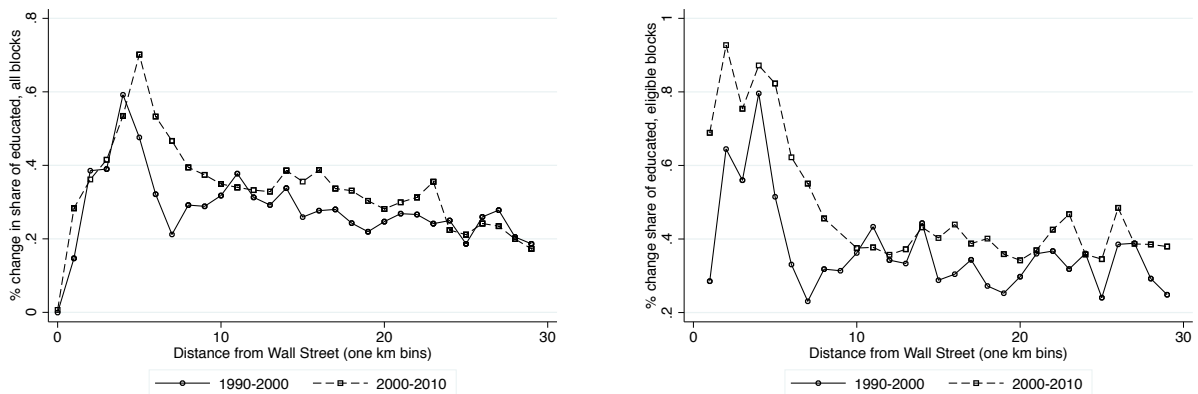
(1) Poor blocks and discrete measure of gentrification,  $\gamma_{it}$ .



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



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



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

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

## Additional references

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

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

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

---

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

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

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

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

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

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