Article

B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2022, Vol. 49(2) 704–721

DOI: 10.1177/23998083211025309
journals.sagepub.com/home/epb
SAGE

Monitoring streets through tweets: Using user-generated geographic information to predict gentrification and displacement

Karen Chapple
University of California at Berkeley, USA

Ate Poorthuis
KU Leuven, Belgium

Matthew Zook
University of Kentucky, USA

Eva Phillips
University of California at Berkeley, USA

Abstract

The new availability of big data sources provides an opportunity to revisit our ability to predict neighborhood change. This article explores how data on urban activity patterns, specifically, geotagged tweets, improve the understanding of one type of neighborhood change—gentrification—by identifying dynamic connections between neighborhoods and across scales. We first develop a typology of neighborhood change and risk of gentrification from 1990 to 2015 for the San Francisco Bay Area based on conventional demographic data from the Census. Then, we use multivariate regression to analyze geotagged tweets from 2012 to 2015, finding that outsiders are significantly more likely to visit neighborhoods currently undergoing gentrification. Using the factors that best predict gentrification, we identify a subset of neighborhoods that Twitter-based activity suggests are at risk for gentrification over the short term—but are not identified by analysis with traditional census data. The findings suggest that combining Census and social media data can provide new insights on gentrification such as augmenting our ability to identify that processes of change are underway. This blended approach, using Census and big data, can

Corresponding author:
Karen Chapple, University of California, 228 Wurster Hall, Berkeley, CA 94720-1850, USA.
Email: chapple@berkeley.edu

help policymakers implement and target policies that preserve housing affordability and protect tenants more effectively.

Keywords
Gentrification, social media, big data, Twitter, neighborhood change

Introduction

Beginning with the Chicago School, generations of urban theorists have studied neighborhood change, whether focused on segregation, gentrification, urban sprawl, or other processes (for example, see Jargowsky, 1997; Lees et al., 2016; Massey and Denton, 1993; Park, 1925). For many decades, studies have either relied primarily on demographic (census) data aggregated at the neighborhood level which mask complex and micro-scale causal dynamics, or utilized in-depth case studies that limit generalization. While new methodological approaches for analysis have been developed (Delmelle, 2015), there remains ample opportunity for innovation. This is particularly the case for attempts to predict future change, particularly gentrification, as efforts to project change based on previous patterns have fallen surprisingly short (Chapple and Zuk, 2016).

The availability of powerful computational approaches and new data sources provides an opportunity to revisit traditional analyses of neighborhood change and increase predictive power. For instance, researchers have used machine learning techniques on existing census data records to analyze existing patterns of neighborhood ascent and decline in order to predict gentrification (Reades et al., 2019). Researchers have also used user-generated data on activity patterns, such as social media data, to refine traditional conceptions of residential segregation or predict housing price changes (Shelton et al., 2015; Steentoft et al., 2018). These types of data are particularly useful since they are available in "real-time" rather than on the periodic schedule of a census. Relatedly, data from other non-Census sources allow for insightful analysis on neighborhood change resulting from new practices such as short-term rentals (Wachsmuth and Weisler, 2018).

Following this vein, this paper explores how data on urban activity patterns, specifically, geotagged tweets, improve the understanding of one type of neighborhood change—gentrification—by identifying dynamic connections between neighborhoods. For this paper, we first develop a typology of neighborhood change from 1990 to 2015 for the San Francisco Bay Area, based on conventional demographic data, and identify non-gentrified areas vulnerable to gentrification, based on risk factors from previous decades. However, this risk category is too broad, both temporally and spatially, to be useful for policymakers seeking to target interventions. In order to narrow this category, we use a blended approach that combines Census data with a new big data source. More specifically, we validate the typology and refine the identification of at-risk areas with more recent data, by analyzing geotagged tweets from 2012 to 2015, to measure the extent to which outsiders (users not living in the tract from which they tweeted) spend time in different types of neighborhoods (see Shelton et al., 2015). Controlling for neighborhood and built environment variables, we identify the factors associated with outsider tweeting during visits and find that outsiders are significantly more likely to visit neighborhoods currently undergoing gentrification. We then determine the factors that best predict recent ongoing gentrification, finding that tweets from outsiders, among other factors, are statistically significant. Using these factors, we

identify a subset of neighborhoods that Twitter-based activity suggests are at risk for gentrification over the short term—but are not identified by analysis with traditional census data. We conclude by discussing the policy implications and implementation possibilities. Combining census data with crowd-sourced big data for urban analysis remains relatively new (see Törnberg and Chiappini, 2020), and blended gentrification studies typically focus on the urban core rather than entire metropolitan regions (Gibbons et al., 2018a; Hristova et al., 2016); thus a key aspect of this paper is understanding how blended data approaches can provide new insights on gentrification as it unfolds over decades across different types of neighborhoods.

Empirical understandings of gentrification: From conventional to big data

Considerable disagreement remains about how to operationalize neighborhood change, particularly gentrification. Most agree that gentrification is a form of neighborhood transformation that involves ascent rather than decline. However, some emphasize flows of capital over people and see gentrification as primarily about disinvestment and devaluation processes that then create a "rent gap" and thus new capital accumulation (Smith, 1979). Others pay more attention to the flows of people and analyze population shifts due to the transformation of production and/or consumption (Ley, 1996; Zukin, 1982). A similar debate has arisen about commercial or cultural gentrification and whether it arises from shifts in cultural consumption or forces of capital accumulation (Hackworth and Rekers, 2005). Yet there is relative consensus that it manifests in retail upscaling that attracts a new class and type of consumer to traditional commercial streets, in a visible and sudden transformation (Zukin et al., 2009).

Researchers often focus on low-income neighborhoods (generally operationalized as census tracts), as these areas are seen as particularly vulnerable to the influx of capital or population (see, for instance, Ding et al., 2016; Ellen and O'Regan, 2011; Freeman, 2005). The global gentrification literature expands gentrification to a wide variety of contexts, actors, and processes, in general de-emphasizing the role of individual consumption (Lees et al., 2016). Studies of U.S. housing markets place particular importance on understanding flows of people rather than capital, making displacement, or forced moves that occur despite having met legal conditions of occupancy, a key focus.

Gentrification-induced displacement may be direct, as landlords evict tenants, or indirect, caused by increasing rents. It may also be exclusionary, taking place when a household is simply unable to move into a neighborhood, typically because of housing costs (Marcuse, 1986). Researchers rarely study displacement via census data, given lack of data about mobility and choice. For instance, in the face of rent increases, some residents must leave, but others choose to stay and pay more rent. Given these challenges, most researchers tend to use simple mobility rates as a proxy for displacement (Ding et al., 2016; Ellen and O'Regan, 2011).

Despite these issues, several studies attempt to create either typologies of gentrification and displacement risk, or stage measures of gentrification, for cities and/or regions using publicly available census data (Atkinson et al., 2011; Bates, 2013; Gibbons et al., 2018a; Regional Plan Association, 2017).[1] Most of these approaches avoid statistical analysis, and none measure displacement empirically, i.e., the number of people leaving a neighborhood. Instead, they assume that hot housing markets result in displacement. Researchers have

generally not bothered to validate their own predictions; one study that did found that it had a 79% false positive rate in predicting gentrification (Chapple and Zuk, 2016).

In this article, we begin by essentially replicating previous approaches, but we use regression to identify risk while also describing a variety of trajectories of neighborhood change in diverse areas throughout the region over an extended (25 year) time period. Given that neighborhood change processes take time to unfold (Chapple and Loukaitou-Sideris, 2019), there is theoretical foundation for using this time-span. However, this fails to capture more rapid or recent changes, including reversals of neighborhood fortune or catalytic changes due to new development.

To counter this temporal disadvantage, this paper turns to a different data source, geolocated social media. User-generated geographic information offers the potential not only to provide more current—even real-time—data, but also allows researchers to pinpoint activity patterns at finer geographies than census tracts, which are not always aligned with "real" neighborhood boundaries (Shelton and Poorthuis, 2019). Moreover, because much of these data are available at a relatively precise "point" scale, researchers can create areal units that match local understandings and specificities. By using data generated by Twitter users, rather than relying on measures of housing prices and tract-level demographics, researchers can also broaden the investigation of neighborhood change to the entire city, rather than just residential areas.

Prior research on neighborhood change using social media data has generally taken one of three tracks. First are studies that analyze perceptions of neighborhood change via social media content such as Zukin et al.'s (2017) analysis of Yelp reviews documenting "discursive redlining" in positive reviews for White-gentrifying neighborhoods, or Boy and Uitermark's (2017) study that shows that Instagram users elevate marquee locations and make others essentially invisible. Second is work that links spatial patterns such as gentrification to social network interactions, typically via blending census and social media data (Ye and Liu, 2018). This includes documenting the connection between check-ins from networks of affluent visitors in deprived neighborhoods and their subsequent economic improvement (Hristova et al., 2016), and the association of gentrifying neighborhoods with interactions among social media users (Gibbons et al., 2018b). These latter two studies also connect neighborhood change explicitly to the rise of popular restaurants, bars, and cafes. Third are projects that use tweets to track activity patterns with a focus on the relationship to housing price changes. Just examining activity spaces of different socio-economic groups via geotagged tweets can develop new conceptions of how neighborhoods are produced (Shelton et al., 2015) and even yield early predictions of rent increases (Steentoft et al., 2018). This paper builds upon and connects elements of the second and third tracks, by tracking the mobility of social media users and using this to create a network between places. This allows us to identify the socio-economic characteristics of outsiders to gentrifying neighborhoods and consider how this new information can help identify areas at risk of change.

To be clear, however, the use of user-generated data does come with both technical and ethical problems. For example, relying on Twitter for its time-stamped geotags has its shortcomings, raising questions of internal validity (Offenhuber, 2017) and creating problems for interpretation. Local context often shapes the choice whether to post on social media, rendering some places or times more likely to be documented than others (Boy and Uitermark, 2017). As a result, the distribution of Twitter activity cannot be assumed to be representative as uneven access, inhospitable contexts, or multiple other factors intervene. Therefore, throughout this paper, we make no claim or use of Twitter data as a representative sample of the population. Instead, we argue that analyzing changes in the spatial footprints of the Twitter platform provides insight to a subset of gentrification processes that augment an understanding based solely on census data. In short, Twitter data does not

allow us to make claims on the general prevalence of gentrification (as we can with Census data). It does provide us with evidence that gentrification-related activity (dynamic connections between places within an urban region) exists, which is why *combining* Census and big data is potentially useful.

Relatedly, great care should be taken in interpretation of the meaning of any social media posting (Poorthuis et al., 2020). For this reason, this paper treats a tweet simply as an indicator of presence rather than ascribing deeper meaning. Ethical concerns are important to consider as gathering larger amounts of social media data allows researchers to build detailed profiles on individual movement and speech (Zook et al., 2017). For this reason, all data presented in this paper are shown at aggregate levels without identifying details. Finally, we recognize that online activity can have a direct effect on how material places function (Zook and Graham, 2007) and can contribute to creating and expanding social divides. However, establishing such a causal connection is challenging and would require a different approach (see Spangler, 2020; Törnberg and Chiappini, 2020) that is tangential to this paper's primary goal of examining how blending census and big data sources can provide new insights on the gentrification process.

Data and methods

This paper requires multiple steps of data preparation. First, we create a standard typology of gentrification using Census data to categorize neighborhoods by stage of gentrification, displacement, and exclusion, generally following methodologies put forth by Bates (2013) and the Urban Displacement Project (Thomas et al., 2020), and definitions of gentrification that include both influxes of new residents and capital, following Freeman (2005) and Ding et al. (2016). Then, to address some of the shortcomings of this conventional approach, most specifically the temporal lag that prevents identifying more recent changes, we examine mobility between neighborhoods as measured by Twitter users.

While we could conduct this analysis for any large metropolitan area, we chose the case of the San Francisco Bay Area megaregion given the prevalence of and rapid change in gentrification throughout the region (Chapple and Zuk, 2016). For the last 50 years, the success (and occasional failures) of the information technology sector has shaped the story of this region: innovative firms have attracted high-skilled young in-migrants, who drive up housing prices in the core and force lower-wage workers into long commutes from throughout the megaregion (Storper et al., 2015). From 2000 to 2015, housing prices in the core of the region doubled, compared to a 75% increase overall in 20 major metropolitan areas in the United States (S&P Dow Jones Indices LLC, 2021). The megaregion includes the five core counties that comprise the metropolitan areas of San Francisco, San Jose, and Oakland (San Francisco, San Mateo, Santa Clara, Alameda, and Contra Costa); the remaining four counties considered part of the commuter shed (Marin, Sonoma, Solano, and Napa); plus Santa Cruz, Sacramento, Yolo, and San Joaquin counties. These counties range in density from urban (18,562 persons per square mile in San Francisco) to largely rural (e.g., 162 persons per square mile in Solano County). Thus, although core areas of the region may be considered an extreme case study in terms of gentrification, the entire study area represents a diversity of urban forms and housing submarkets (Yin, 2017).

Creating a typology of neighborhood change

In order to designate neighborhoods by their stage of change, we use data from the Decennial Census from 1990 and 2000, and American Community Survey for the years

2011–2015. To reconcile the changes in tract boundaries from earlier time periods, we used Brown University's Longitudinal Tract Data Base (LTDB), which normalizes census tract data from each year to 2010 census tract boundaries to maximize comparability across the study period. In the case of variables not provided by the LTDB, we downloaded the original raw data and used LTDB's crosswalk.

We characterize change in both low- and high-income neighborhoods, looking at gentrification and displacement in the former and exclusion in the latter. We thus divide the region into low-income neighborhoods at 80% of area (county) median household income or less, and moderate- to high-income neighborhoods with median income above 80% of area (county) median household income. We select these thresholds to be consistent with affordable housing policies and programs. (Table S1 presents the full typology methodology.)

To describe the neighborhoods where gentrification and displacement are taking place, most studies first pinpoint the neighborhoods with potential to change (or the "eligible" tracts) (Chapple and Zuk, 2016; Ding et al., 2016; Freeman, 2005; Gibbons et al., 2018b). To do this, we use several different indicators. Vulnerability to change is defined by availability of affordable housing, so we select neighborhoods where either rents or housing values are below the area (county) median. We also select areas with low college education, and high low-income households, renters, and nonwhite households, all relative to the median. "Eligible" neighborhoods must have housing affordability plus any two of the four demographic characteristics.

We characterize neighborhood change in the form of *gentrification* (operationalized as the influx of investment and people into low-income areas); *displacement* (the loss of low-income households without replacement, in low-income areas), and *exclusion* (when displacement is occurring in high-income neighborhoods). We measure gentrification via the change in real median housing value or rent above the county median change, as well as growth in share of college-educated population and household median income greater than the median change; we also include a measure of exclusionary displacement, i.e., decrease in the in-migration rate of low-income households. To measure displacement and exclusion, we use two indicators—absolute loss of low-income households between census years and decrease in in-migration.

We rely on logit regression to identify the key long-term predictors of ongoing gentrification in eligible (vulnerable) tracts, analyzing the entire 25-year period from 1990 to 2015 (Table 1). For gentrification, the significant factors include high shares of historic housing (built pre-1950) and recent new housing construction; a central city location; and lack of local housing and employment density, as well as households with children.[2] These are the factors typically identified by the literature, with the exception of the density variables. This unintuitive finding likely results from the diverse urbanization patterns in the 13 counties under study: for example, high-density tracts may either have very little housing or be dominated by high-end apartment buildings with little opportunity for gentrification, while rural towns may have gentrifying cores surrounded by low employment and population density. We categorize neighborhoods that score above the regional median (or below, for the negative factors) on at least half the significant factors as at risk.

Figure 1 maps the results of this classification, both identifying historic change and predicting areas likely to change in the future. Overall, of the 2138 census tracts, 5% have already gentrified, 10% are currently undergoing gentrification, 3.5% are currently undergoing displacement without gentrification, 31% are in some stage of exclusion, and 11% are at risk of gentrification (see Table S1). The remaining 39% of tracts are classified as stable.

Table 1. Factors predicting ongoing gentrification, 1990–2015.

Type	Variable	Coefficient	Odds ratio
Socio-economic characteristics	Household income, 1990	–0.000[b]	1.000
	Household income, 1990, squared	0.000[a]	1.000
	% of households with children, 1990	–5.566[a]	0.004
	% of Latinx residents, 1990	1.268	3.554
	% of African-American residents, 1990	1.149	3.155
	% of Asian residents, 1990	–1.214	0.297
	% of college-educated residents, 1990	–0.262	0.769
	% of renters, 1990	3.312[a]	27.444
Built environment characteristics	Central city location (dummy)	0.977[a]	2.656
	Transit neighborhood, 1990 (dummy)	0.032	1.033
	Transit neighborhood, 1990–2000 (dummy)	–0.008	0.993
	Population density, 1990	–0.000[c]	1.000
	Employment density, 1990	–0.000[c]	1.000
	% of housing units built before 1950	3.136[a]	23.005
	% of housing units built 1980–1990 (recent)	1.991[b]	7.326
	% of residents commuting by car, 1990	–0.266	0.024
	Constant	–0.439	0.645

[a] $p = 0.000$. [b] $p < 0.05$.
[c] $p < 0.10$.
Notes: $N = 386$; % correctly predicted 77.7%; pseudo R-square = 0.342; $\chi^2 = 0.000$, –2 log likelihood = 361.015.

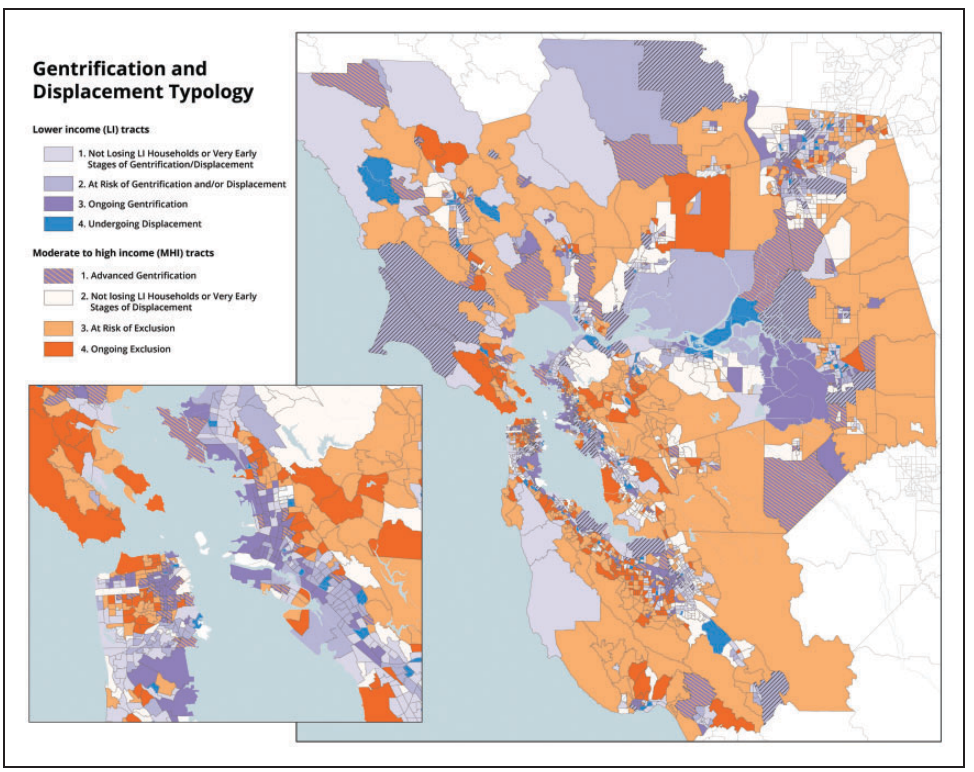

Figure 1. Typology of gentrification and displacement in the San Francisco Bay Area.

While not discounting the work required to produce this typology or its usefulness, one potential critique is that it identifies a relatively large number of "at-risk" neighborhoods (or census tracts), 241 in total. Given limited resources for policy making, an important question is how this prediction might be refined, particularly controlling for more recent activity than available from the census. For this, we now turn to the second stage of the analysis to use geo-tagged Twitter data to investigate how it helps distinguish neighborhood types and identify areas where change is imminent.

Preparing Twitter data. We began our analysis with a dataset of geotagged tweets (54.5 million in total) for the 13 counties of the San Francisco Bay Area megaregion sent between July 2012 and June 2015, collected by the DOLLY archive at the University of Kentucky. In order to link this data to our gentrification typology, we followed five steps: determining home and adjacent location, eliminating "power" users, identifying characteristics of the home tract, and adding new control variables.

Identifying home and adjacent location. To determine the home location for users, we first removed all users with less than 20 tweets in our dataset to ensure a sufficient number of observations. The remaining 51.8 million geotagged tweets were sent by 183,715 users. For these, we use a relatively simple filtering algorithm to determine the most likely home location, or more specifically the location where users had a sustained presence. A home location (defined as a census tract) needs to be the location of at least 10 tweets, sent on at least seven different days and during eight or more different hours of the day. If multiple locations satisfy, we use the location with the most tweets. With this method, we are able to assign a home location for 102,338 users who sent a total of 47.65 million tweets within a census tract location.[3] This represents only 13.5% of all users, and 55.7% of those users who sent more than 20 tweets during the study period. Although different values could be chosen for the parameters used in the inference of the most likely home location, we opt for a relatively conservative approach to prevent assigning users to a location that neither is their home nor plays any role as a "base" location in their life. In this sense, we have designed this procedure to prevent false positives (cf. Chen and Poorthuis, 2021 for a sensitivity analysis and discussion of different home location algorithms).

Using nearest neighbor analysis in ArcGIS, we identify tracts adjacent to the home tract, herein called "neighbor" tracts. From this we categorize all remaining tweets by user-type: local, neighbor, or non-local (outsider). These categories of user-type, relative to the census tract, are defined as follows:

- Local: User's home location is within the census tract.
- Neighbor: User's home location is adjacent to the census tract.
- Non-local (outsider): User's home location is neither in nor adjacent to the census tract.

The neighbor category allows us to differentiate between local and non-local outsiders and address boundary issues: for instance, tweets that emanate from opposite sides of the street, from two different census tracts, might otherwise be categorized as home and outsider, but are both essentially local. Figure 2 illustrates the three tweet types in the City of San Francisco, using only 1% sample of the tweets for visual clarity. The figure validates our approach by suggesting a concentration of outsider tweets not only in the downtown central business district, but along the city's major arterials. We acknowledge the imprecision of the "outsider" definition, i.e., outsiders are not necessarily indicative of gentrification and could represent people going to work or visiting family and friends. Nevertheless, the

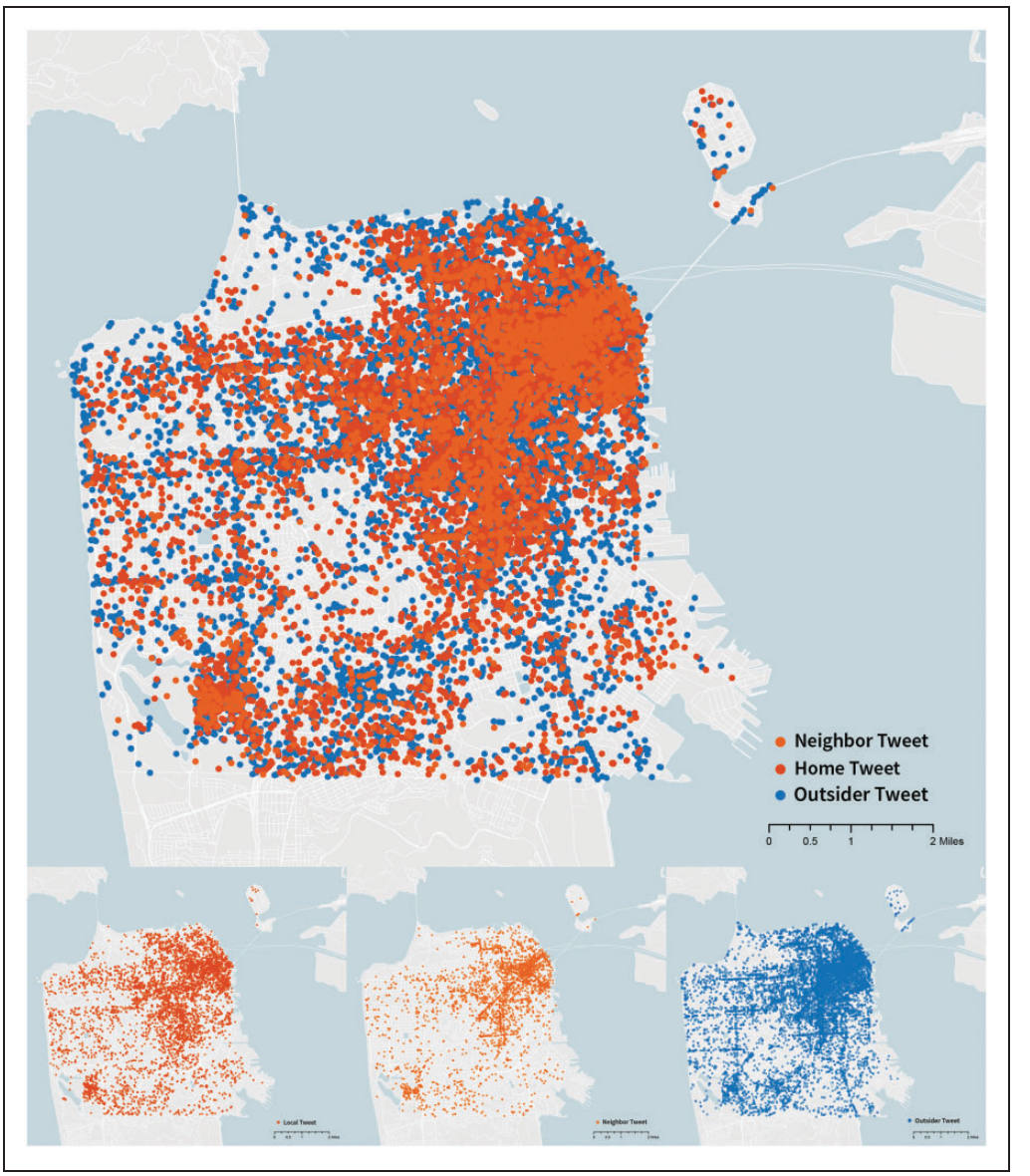

Figure 2. Geotagged tweets in the City of San Francisco, 2012–2015 (1% sample).

value in its use is that it captures a dynamic that is absent from the residential census (movement between places). Despite potential imperfections in this operationalization, exploring what new insights may be gained is a worthwhile endeavor. Moreover, in our regression models, we control for commercial land use in an attempt to account for the possibility that this indicator is capturing work locations.

Eliminating power users. It is also important to control for extremely active users, or power-users, so that the activity of a single Twitter account does not unduly shape the results by tweeting repeatedly from the same location. To do this, we filter the data to

include only one tweet per user per day per tract, which still gives frequent visitors relative importance in the dataset, and leaves 14,585,347 tweets and 102,338 users in the dataset.[4]

Identifying characteristics of the home tract. Given our ultimate goal of understanding the nature of visitors to gentrifying neighborhoods, we need to assign users demographic characteristics. We do this by associating users (and their tweets) the characteristics, including income, race/ethnicity, and education, of their home census tract according to the 2011–2015 American Community Survey. Income is defined relative to the county median household income: over 120% and under 80% of county median qualified as high-income and low-income respectively, and the remaining as middle-income. User with home tracts with concentrations of racial/ethnic groups over the county median are characterized as dominant Latinx, non-Hispanic Black, or Asian. Likewise, we define users as coming from college-educated tracts if the percentage of college-educated residents in their home tract is greater than the county median. Because of the potential ecological fallacy that results from this approach, we use these socio-economic characteristics not in the models but in exploratory descriptive statistics.

Other characteristics. To characterize the built environment where tweeting takes place, we use variables measuring central city location (San Jose, San Francisco, or Oakland); transit neighborhood (location in a census tract that intersects the one-half-mile radius of a fixed rail transit station); percent of tract parcels in commercial land use (calculated using county tax assessor data); employment density (calculated using the Longitudinal Employer-Household Dynamics data); population density (calculated from the American Community Survey); and housing age (calculated from the American Community Survey). These variables proxy for the elements of centrality, accessibility, and architectural quality described as driving gentrification in the literature, as well as the connection between commercial uses, social media use, and neighborhood change (Gibbons et al., 2018b; Hristova et al., 2016).

We begin the analysis by describing the characteristics of users (based on their home tracts) tweeting in different neighborhoods and link this to our typology of gentrification and displacement. Next, we examine how outsider tweets improve our gentrification prediction model in Table 1. Then, we use ordinary least squares (OLS) regression to identify the factors behind outsider tweets in a particular neighborhood, with a close look at how neighborhood types predict visits from these outsiders. After establishing the relationship between outsiders and neighborhoods undergoing gentrification, we conclude by using tweeting patterns to help refine the identification of neighborhoods at risk for gentrification.

Analyzing tweeting by neighborhood and user type

To determine tweet patterns by neighborhood, we examine how tweeting by the three types of users—local, neighbor, and outsider—varies across neighborhoods defined by socio-economic and built environment characteristics. The average tract had 2,409 geotagged tweets from local users, 810 tweets from users in neighboring tracts, and 3,292 tweets from outsiders. In general, locals and neighbors tweet about the same amount regardless of neighborhood type, although users from tracts with above-median shares of African-American and Asian residents, as well as very high-income residents, are slightly more likely to tweet from their local tract (Table S2 provides location quotients based on average tweets by user type in each neighborhood type). Concentrations of outsider tweets occur across a variety of neighborhood types including those lacking: households with children, very-high

income households, and Latinx residents. Outsider tweets are also concentrated in tracts with higher than median levels of older millennials (age 25–34), Asians, college-educated, and very low-income residents. Tracts located in the central city or transit neighborhoods, with a high share of commercial parcels, high employment and population densities, and concentrations of both new and historic housing are also characterized by high location quotients for outsider tweets.

There is also a correlation between outsider tweets and gentrification, with very high concentrations in neighborhoods undergoing gentrification (LQ 1.81) or classified as experiencing advanced gentrification (LQ 1.32). While gentrifying neighborhoods also have high concentrations of local Twitter users, this is also observed in stable moderate to high-income neighborhoods albeit without the same level of outsider activity.

Given the concentration of outsider tweeting in gentrifying neighborhoods, it is important to understand where these outsiders are coming from. A disproportionate share comes from people associated with neighborhoods that are also either undergoing gentrification or in a state of advanced gentrification (Figure S1).

The distribution of outsider tweeting across the region is uneven not only in terms of neighborhood types but also according to the home tract demographics associated with users. Outside users from low-income tracts tweet disproportionately in the East Bay, while users from middle-income tracts tweet in San Francisco and the South Bay (Figures S2 and S3). Users from tracts with higher education levels tend to tweet in the urban core and university campuses (Figures S4 and S5). Users from both African-American and Latinx neighborhoods tend to tweet in the East Bay (Figures S6 and S7). More generally speaking, neighborhoods with disproportionate outside tweeting tend either to be gentrifying or visitor locations, such as Golden Gate Park, airports, or downtowns.

Another lens into tweeting patterns comes from the timing of tweets. In neighborhoods with ongoing or advanced gentrification, there is a disproportionate share of outsiders tweeting across all time periods (Figure S8).

Using outsider tweets to predict gentrification

In this section, we shift to multivariate analysis to better understand the relationship between outsider tweeting and gentrification. The census data used to create the gentrification typology (1990, 2000, and 2011–2015) mostly predates our Twitter dataset (2012–2015), meaning that outsider tweeting may represent either a predictor of gentrification or an outcome of gentrification. To account for this, we model gentrification as both a dependent and an independent variable. More specifically, we examine associations with outsider tweeting during visits. Finally, we expand our original model of factors predicting gentrification over a 25-year period to include outsider tweets.

What, then, is associated with disproportionate tweeting from outsiders in a neighborhood? Using the independent variables used earlier in creating location quotients (see Table S2), we use OLS regression to predict the number of outsider tweets in each tract resulting in an adjusted R^2 of 0.28 and no issues of multi-collinearity or spatial autocorrelation (see Table 2). The regression identifies a number of socio-economic and built environment characteristics that predict outsider tweets. These include: a higher share of college-educated and millennial (25–34 year old) residents; a lower share of households with children predict outsider tweets; locations within the central city or a transit neighborhood; high shares of commercial parcels and historic housing; and low densities both in terms of employment and population. Most significantly for this project, neighborhoods categorized as low-income in which ongoing gentrification is taking place were also a significant

Table 2. Predicting outsider tweets.

Type	Variable	B	Beta	t-stat
Twitter	Total number of tweets, 2012–2015	0.000	0.094[a]	0.001
Socio-economic characteristics	% of households with children, 2015	–0.183	–0.154[a]	0.000
	% of residents aged 25–34 years, 2015	0.088	0.043	0.130
	% of Latinx residents, 2015	–0.023	–0.027	0.492
	% of African-American residents, 2015	–0.004	–0.008	0.745
	% of Asian residents, 2015	0.005	0.035	0.150
	% of college-educated residents, 2015	0.101	0.143[a]	3.544
	Central city location (dummy)	0.024	0.074[a]	0.005
	Transit neighborhood (dummy)	0.034	0.105[a]	0.000
Built environment characteristics	% of parcels in commercial land use	0.717	0.319[a]	0.000
	Employment density, 2014	–0.000	–0.079[b]	0.009
	Population density, 2015	–0.000	–0.183[a]	0.000
	% of housing units built before 1950	0.062	0.101[a]	0.001
	% of housing units built after 2000	0.024	0.021	0.385
Neighborhood type	Low-income at risk	–0.007	–0.014	0.572
	Low-income ongoing displacement	0.029	0.036	0.113
	Low-income ongoing gentrification	0.032	0.064[b]	0.011
	Advanced gentrification	–0.006	–0.009	0.715
	Moderate/high-income at risk for exclusion	–0.013	–0.034	0.238
	Moderate/high-income ongoing exclusion	–0.007	–0.014	0.591
	Constant	0.425[a]	19.693	
	Adj. R^2	0.282		
	Significance	0.000		
	N	2138		

[a] $p = 0.000$.
[b] $p < 0.05$.
[c] $p < 0.10$.

predictor of outsider tweets, in fact the only neighborhood type that was significant. This suggests then that this type of neighborhood change is indeed distinct from others, including ongoing displacement, where residents are moving out but often in a context of disinvestment instead of gentrification.

Shifting to using outsider tweets as an independent variable, Table 3 revisits the regression presented in Table 1, with the inclusion of outsider tweets. The effect of outsider tweets is significant, and its impact only trails that of share of renters and historic housing. Adding the variable improves the model slightly, most notably for gentrifying tracts, predicting 79.8% of the cases (54% of the gentrified cases) rather than 77.7% (48% of the gentrified).

This association between visitors tweeting in a tract and gentrification suggests that outsider tweets could contribute to an early warning system for gentrification. In the next section, we explore what that could look like.

Identifying tracts at risk

The neighborhood change typology identified eight risk factors for gentrification and found that, based upon current characteristics, some 241 census tracts were at risk for change over the long-term. Using data on outsider tweets, we can identify which of these tracts are already experiencing heightened activity from outsiders. Of the at-risk tracts, 108 are

Table 3. Factors predicting ongoing gentrification 1990–2015, including outsider tweets.

Type	Variable	Coefficient	Odds ratio
Socio-economic characteristics	Household income, 1990	–0.000[a]	1.000
	Household income, 1990, squared	0.000[a]	1.000
	% of households with children, 1990	–4.928[b]	0.007
	% of Latinx residents, 1990	1.749	5.751
	% of African-American residents, 1990	1.851	6.366
	% of Asian residents, 1990	–1.057	0.348
	% of college-educated residents, 1990	0.590	1.804
	% of renters, 1990	3.209[a]	24.748
Built environment characteristics	Central city location (dummy)	0.852[b]	2.345
	Transit neighborhood, 1990 (dummy)	–0.064	0.938
	Transit neighborhood, 1990–2000 (dummy)	–0.127	0.881
	Population density, 1990	–0.000	1.000
	Employment density, 1990	–0.000[c]	1.000
	% of housing units built before 1950	3.166[a]	23.719
	% of housing units built 1980–1990 (recent)	1.996[b]	7.358
	% of residents commuting by car, 1990	0.631	1.880
	% of tweets from outsiders	3.010[a]	20.295
	Constant	–3.301	0.370

[a] $p = 0.000$.
[b] $p < 0.05$.
[c] $p < 0.10$.
Notes: $N = 386$; % correctly predicted 79.8%; pseudo R-square = 0.364; $\chi^2 = 0.000$; –2 log likelihood = 352.972.

currently experiencing an above-median share of outsider tweets. Figure 3 shows the number of traditional risk factors associated with each tract, and adds cross-hatching to identify which are also experiencing an above-median outsider Twitter activity. Although areas throughout the region's cities and suburbs are at risk, the areas with disproportionately high incidence of outsider tweets tend to be either adjacent to gentrified areas (e.g., Oakland, San Leandro, south San Francisco, Sacramento), tourist areas (e.g., Point Reyes), or college towns (e.g., Santa Cruz). Outsider tweets are also correlated with several of the same risks identified in the neighborhood change typology, including low share of households with children, high share of historic housing, central city location, and high share of renters (see Table S3). However, some other risk factors are not correlated with this Twitter activity, most notably a hot real estate market. This is consistent with the idea that outsider tweets may act as an early indicator of gentrification activity, before real estate activity has increased significantly.

In an effort to develop a more refined understanding of gentrification risk and its association with outsider tweets, we compare the characteristics of neighborhoods at risk to already gentrifying neighborhoods. If low-income at-risk tracts share similar attributes with gentrifying areas (including the disproportionate presence of outsider tweets), they might be particularly susceptible. We explore this by conducting difference of proportions tests to examine the differences in a set of characteristics in at risk and gentrifying neighborhoods with below or above median share of outsider tweets. Tracts with ongoing gentrification share certain demographic characteristics (high shares of college-educated residents and 25–34 year-olds, low shares of Latinx residents), as well as certain built environment characteristics (walkability, transit accessibility, commercial land use, and density) (Table S4). Outside Twitter users coming to these neighborhoods come disproportionately from high-educated, high-income, and non-Latinx neighborhoods.

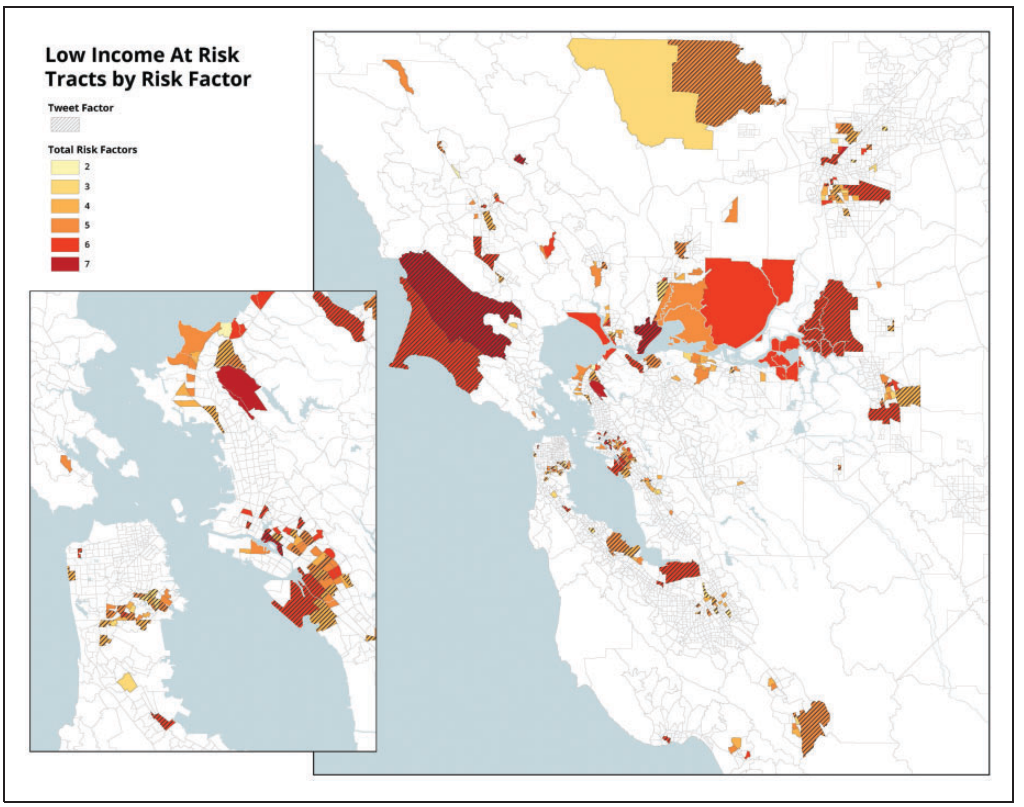

Figure 3. Tracts at risk for gentrification by risk factor and outsider tweets.

Using this information, we might further refine our subset of 108 low-income neighborhoods at risk. For example, commercial neighborhoods near transit with a concentration of 25–34 year-olds within this group of 108 include Broadway Auto Row in Oakland and downtown Redwood City. Other neighborhoods that are highlighted via this exercise include Portola in San Francisco (a walkable, dense neighborhood with a relative low share of Latinx residents) and El Sobrante in Contra Costa County, a diverse community with low levels of college education and housing appreciation. Other neighborhoods like Bolinas (Marin County) are surprises given their existing concentrations of high-income, high-educated, non-Latinx users living alongside long-term low-income residents, but do experience high levels of tourism that could put oldtimers at risk.

Conclusion and policy implications

Neighborhood change leads to windfall profits for some and displacement for others. Stakeholders from real estate speculators to policy makers seek the ability to predict how change will transpire. This research first analyzes the performance of conventional methods of analyzing potential risk of gentrification, and then shows how Twitter data can not only improve prediction but also help pinpoint areas at risk even when real estate activity has not yet accelerated. This study augments previous studies pioneering blended approaches (such as Gibbons et al., 2018b and Hristova et al., 2016) in several ways: identifying home locations of Twitter users, examining the socio-economic characteristics of visitors to gentrifying areas,

blending Twitter data with a typology of neighborhood change, controlling for commercial land use, and adopting a study area that encompasses an entire, diverse megaregion.

The findings suggest that social media data, blended with census data, can serve as an early warning indicator that processes of change are underway. When outsiders tweet in a neighborhood, it is more likely to be undergoing gentrification, controlling for other factors such as commercial land use. The presence of tweets from outsiders in a low-income neighborhood that has not yet gentrified may suggest that the area is about to undergo change. Certain types of areas may be more susceptible, such as higher-density, more walkable commercial neighborhoods near transit. Tweets that are disproportionately from people who live in higher-income, higher-educated, less ethnically diverse neighborhoods may be an indication of change to come.

This study has certain limitations that future research can hopefully address. First, because we only had three years of geotagged tweets, we were not able to use this data to look at change over time, instead linking it to census data from different timeframes. Future studies should attempt to link time periods more precisely. Second, ascribing home neighborhood demographic characteristics to Twitter users is an imperfect proxy given the diversity of neighborhood residents; still, to the extent that it is possible to identify user demographics more precisely, this could be a fertile area for further research. Third, these findings rely on just one form of social media data, Twitter; yet other forms with growing popularity, such as Instagram, may prove to be even more powerful in predicting neighborhood change. Fourth, areas undergoing (or about to undergo) gentrification are likely experiencing population turnover generally, so qualitative research might help to reveal exactly what mobility data are capturing. Finally, aggregating Twitter data into districts creates statistical bias in the form of the modifiable areal unit problem, which methods such as applying a hexagonal grid can reduce (Openshaw, 1983).

Nonetheless, our findings have important implications for policymakers seeking to mitigate the negative impacts of gentrification, which can uproot longstanding communities. Models that try to predict gentrification often result in false positives, and incorporating new data sources should help produce more accurate predictions. Previous studies focus primarily on dense urban gentrifying areas, while our results encompass both rural and urban areas, which make them relevant across different jurisdictions. A better understanding of where change is about to occur will help policymakers implement and target policies that preserve housing affordability and protect tenants more effectively. Of course, private sector real estate interests are actively using big data to do the same (Stewart, 2019), but tools like these can empower the public sector to keep pace.

Acknowledgements

The authors would like to thank Catherine Bui and Kush Khanolkar for research assistance, and the anonymous referees for their insightful comments.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

Notes

1. Notably, other gentrification researchers have augmented Census data with data from other sources—such changes to the built environment (Wyly and Hammel, 1998) and mortgage borrowing (Wyly and Hammel, 1999).
2. Household income, income squared, and renter are also significant, but are measured in the dependent variable and thus endogenous, so we do not consider them risk factors.
3. Of the 2250 total census tracts, only 314 slightly underrepresented tweets with home location (less than 0.03% missing) relative to all tweets. Thus, the tracts with home location are a good overall representation of the population of geotagged tweets. To eliminate noise from tourists, we dropped all tweets with a home tract outside the megaregion.
4. Another approach might have been to weight the users by number of tweets. We instead limited to one tweet per user per day in order to account for repeat visits. In other words, we are not just interested in measuring the diversity of users, but whether users visited repeatedly.

References

Atkinson R, Wulff M, Reynolds M, et al. (2011) Gentrification and displacement: The household impacts of neighbourhood change. Report, Australian Housing and Urban Research Institute, Australia.

Bates L (2013) Gentrification and displacement study: Implementing an equitable inclusive development strategy in the context of gentrification. Report, Portland State University, Oregon.

Boy JD and Uitermark J (2017) Reassembling the city through Instagram. *Transactions of the Institute of British Geographers* 42(4): 612–624.

Chapple K and Loukaitou-Sideris A (2019) *Transit-Oriented Displacement or Community Dividends: Understanding the Effects of Smarter Growth on Communities*. Cambridge, MA: MIT Press.

Chapple K and Zuk M (2016) Forewarned: The use of neighborhood early warning for gentrification and displacement. *Cityscape* 18(3): 109–130.

Chen Q and Poorthuis A (2021) Identifying home locations in human mobility data: An open-source R package for comparison and reproducibility. *International Journal of Geographical Information Science* 35(7): 1425–48. doi: 10.1080/13658816.2021.1887489.

Delmelle EC (2015) Five decades of neighborhood classifications and their transitions: A comparison of four US cities, 1970–2010. *Applied Geography* 57: 1–11.

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

Ellen IG and O'Regan KM (2011) How low income neighborhoods change: Entry, exit, and enhancement. *Regional Science and Urban Economics* 41(2): 89–97.

Freeman L (2005) Displacement or succession? Residential mobility in gentrifying neighborhoods. *Urban Affairs Review* 40(4): 463–491.

Gibbons J, Barton M and Brault E (2018a) Evaluating gentrification's relation to neighborhood and city health. *PLoS One* 13(11): e0207432.

Gibbons J, Nara A and Appleyard B (2018b) Exploring the imprint of social media networks on neighborhood community through the lens of gentrification. *Environment and Planning B: Urban Analytics and City Science* 45(3): 470–488.

Hackworth J and Rekers J (2005) Ethnic packaging and gentrification: The case of four neighborhoods in Toronto. *Urban Affairs Review* 41(2): 211–236.

Hristova D, Williams MJ, Musolesi M, Panzarasa P, and Mascolo C (2016) Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. *Proceedings of the 25th International Conference on World Wide Web*, 21–30.

Jargowsky PA (1997) *Poverty and Place: Ghettos, Barrios, and the American City*. New York, NY: Russell Sage Foundation.

Lees L, Shin HB, and López-Morales E (2016) *Planetary Gentrification*. Hoboken, NJ: John Wiley and Sons.

Ley D (1996) *The New Middle Class and the Remaking of the Central City*. *Oxford Geographical and Environmental Studies*. Oxford, UK: Oxford University Press.

Marcuse P (1986) Abandonment, gentrification, and displacement: The linkages in New York City. In: Smith N and Williams P (eds) *Gentrification of the City*. London, UK: Routledge, pp.153–177.

Massey DS and Denton NA (1993) *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard University Press.

Offenhuber D (2017) Sticky data: Context and friction in the use of urban data proxies. In: Kitchin R, Lauriault T and McArdle G (eds) *Data and the City*. Abingdon, UK; Oxon, UK: Routledge, pp.118–128.

Openshaw S (1983) *The Modifiable Areal Unit Problem*. Norwick, UK: Geo Books.

Park RE (ed) (1925) *The City: Suggestions of Investigation of Human Behavior in the Urban Environment*. Chicago, IL: University of Chicago Press.

Poorthuis A, Power D and Zook M (2020) Attentional social media: Mapping the spaces and networks of the fashion industry. *Annals of the American Association of Geographers* 110(4): 941–966.

Reades J, De Souza J and Hubbard P (2019) Understanding urban gentrification through machine learning: Predicting neighbourhood change in London. *Urban Studies* 56(5): 922–942.

Regional Plan Association (2017) Pushed out: Housing displacement in an unaffordable region. Report, Regional Plan Association, New York.

S&P Dow Jones Indices LLC (2021) S&P/Case-Shiller 20-City Composite Home Price Index [SPCS20RSA]. FRED, Federal Reserve Bank of St. Louis. Available at: https://fred.stlouisfed.org/series/SPCS20RSA (accessed 9 March 2021).

Shelton T and Poorthuis A (2019) The nature of neighborhoods: Using big data to rethink the geographies of Atlanta's neighborhood planning unit system. *Annals of the American Association of Geographers* 109(5): 1341–1361.

Shelton T, Poorthuis A and Zook M (2015) Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning* 142: 198–211.

Smith N (1979) Toward a theory of gentrification: A back to the city movement by Capital, not people. *Journal of the American Planning Association* 45(4): 538–548.

Spangler I (2020) Hidden value in the platform's platform: Airbnb, displacement, and the unhoming spatialities of emotional labour. *Transactions of the Institute of British Geographers* 45(3): 575–588.

Steentoft AA, Poorthuis A, Lee B, et al. (2018) The canary in the city: Indicator groups as predictors of local rent increases. *EPJ Data Science* 7(1): 7–21.

Stewart M (2019) The real estate sector is using algorithms to work out the best places to gentrify. *Failed Architecture*. Available at: https://failedarchitecture.com/the-extractive-growth-of-artificially-intelligent-real-estate/ (accessed 4 June 2021).

Storper M, Kemeny T, Makarem N, et al. (2015) *The Rise and Fall of Urban Economies: Lessons from San Francisco and Los Angeles*. Palo Alto, CA: Stanford University Press.

Thomas T, Hartmann C, Driscoll A, et al. (2020) *The Urban Displacement Replication Project: A Modified Gentrification and Displacement Methodology for the Atlanta, Chicago, Denver, and Memphis SPARCC Regions*. Berkeley, CA: Urban Displacement Project.

TöRnberg P and Chiappini L (2020) Selling black places on Airbnb: Colonial discourse and the marketing of black communities in New York 'city. *Environment and Planning A: Economy and Space* 52(3): 553–572.

Wachsmuth D and Weisler A (2018) Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space* 50(6): 1147–1170.

Wyly EK and Hammel DJ (1998) Modeling the context and contingency of gentrification. *Journal of Urban Affairs* 20(3): 303–326.

Wyly EK and Hammel DJ (1999) Islands of decay in seas of renewal: Housing policy and the resurgence of gentrification. *Housing Policy Debate* 10(4): 711–771.
Ye X and Liu X (2018) Integrating social networks and spatial analyses of the built environment. *Environment and Planning B: Urban Analytics and City Science* 45(3): 395–399.
Yin RK (2017) *Case Study Research and Applications: Design and Methods*. Thousand Oaks, CA: Sage Publications.
Zook MA and Graham M (2007) The creative reconstruction of the internet: Google and the privatization of cyberspace and DigiPlace. *Geoforum* 38(6): 1322–1343.
Zook M, Barocas S, Boyd D, et al. (2017) Ten simple rules for responsible big data research. *PLOS Computational Biology 13(3): e1005399*.
Zukin S (1982) *Loft Living: Culture and Capital in Urban Change*. Baltimore, MD: Johns Hopkins University Press.
Zukin S, Lindeman S and Hurson L (2017) The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *Journal of Consumer Culture* 17(3): 459–479.
Zukin S, Trujillo V, Frase P, et al. (2009) New retail capital and neighborhood change: Boutiques and gentrification in New York city. *City & Community* 8(1): 47–64.

Karen Chapple is a professor and Chair of City and Regional Planning at the University of California, Berkeley, where she holds the Carmel P. Friesen Chair in Urban Studies. She studies inequalities in the planning, development, and governance of cities and regions in the U.S. and Latin America, with a focus on economic development and housing.

Ate Poorthuis is an assistant professor of Big Data and Human-Environment Systems in the Department of Earth and Environmental Sciences at KU Leuven. His research explores the possibilities and limitations of big data, through quantitative analysis and visualization, to better understand how our cities work.

Matthew Zook is a University Research Professor in the Department of Geography at the University of Kentucky. His research focuses on how digital technologies and big data are changing cities and the spatial economy.

Eva Phillips is an alumna of the UC Berkeley College of Environmental Design and the Urban Displacement Project, now employed by the Knowledge, Impact and Strategy team at Enterprise Community Partners in New York City. Her work focuses on using novel data sources to better understand neighborhood change and building data tools to inform advocacy and policy.