Using social media data to identify neighbourhood change

Alexis Comber[1]*, Minh Le Kieu[2], Bui Thanh[3] and Nick Malleson[1]

[1]School of Geography, University of Leeds, UK
[2]Department of Civil and Environmental Engineering, University of Auckland, New Zealand
[3]Faculty of Geography, VNU University of Science, Viet Nam
*email: a.comber@leeds.ac.uk

Abstract

This paper explore the use social media data from Twitter to capture perceptions of neighbourhood characteristics, in relation to gentrification. It does this by defining a rudimentary lexicon of words associated with gentrification. This was used to calculate gentrification scores for geo-located tweets. These were then interpolated to create the surfaces in to determine spatial and temporal patterns of gentrification. There are a number of limitations to the methods used in this study, which are discussed and a number of related areas of future work are indicated.

If accepted, this paper will be formatted using the AGILE overleaf template.

Introduction

This short paper describes a novel approach for quantifying neighbourhood spatio-temporal dynamics through the integration of social media data (SMD) to be incorporated within hedonic house price models (HHPMs). It uses analyses of SMD content to identify patterns of emergent gentrification pressure and a future analysis will quantify their impact on local housing markets to identify nascent house price bubbles. In this way, the over-arching aim of this research is to support policy and urban planning and to better understand the drivers of neighbourhood change . In particular, the important role that social factors play in house price dynamics, which is a substantial research gap. The research reported in this paper has the objective of quantify the spatial and temporal trends in perceived neighbourhood characteristics from SMD using natural language processing and to examine the strengths and weaknesses of digital footprint datasets like SMD as proxy measures for capturing neighbourhood level perceptions and processes. It takes the first steps to addresses a number of critical research gaps:

1. Unexplored SMD opportunities to understand Neighbourhood processes. Emerging social phenomena at local levels can be effectively captured through SMD across both spatial and temporal dimensions. Among these phenomena, gentrification stands out as a significant driver of housing price dynamics. Despite its importance, there remains a gap in systematically analysing social media data to detect the emergence of gentrification, as has been done with more traditional datasets (Gray, Buckner, and Comber 2021). Nonetheless, SMD has been utilized to investigate various aspects of neighbourhood socio-demographic characteristics (Lansley and Longley 2016), track neighbourhood dynamics and the evolution of social processes, particularly in urban areas (Poorthuis, Shelton, and Zook 2022), and capture neighbourhood-level behaviours and social dynamics (Nguyen et al. 2016). While some qualitative analyses of SMD have demonstrated its efficacy in revealing emergent gentrification (Bronsvoort and Uitermark 2022), previous studies have predominantly approached the analysis of social process dynamics in a qualitative and fragmented manner. Moreover, quantitative analyses that have been conducted often focused solely on spatial characteristics, neglecting spatio-temporal dynamics. This study addresses these gaps by employing a spatio-temporal approach to SMD analysis, specifically investigating the emergence and establishment of neighbourhood gentrification.

2. Social media use as a lens to understand Gentrification. Gentrification is closely linked to

Table 1: Counts of the tweets collected over the study area, for each year, and their distribution across the neighbourhood areas (LSOAs).

Year	Count	Min	Q1	Median	Mean	Q3	Max
2014	801,662	8	270	433	400	545	1,193
2015	776,790	1	157	287	391	497	3,300
2016	363,193	1	35	83	183	184	4,455
2017	331,532	1	19	56	168	156	4,999
2018	263,360	1	15	46	136	126	4,810
2019	207,334	1	9	31	108	95	4,727

increased usage of social media platforms (Gibbons, Nara, and Appleyard 2018; Walters and Smith 2024; Bronsvoort and Uitermark 2022), although it manifests diversely. Qualitative investigations into individuals involved in gentrification consistently reveal high levels of social media engagement, albeit for varying reasons. Bronsvoort and Uitermark (2022) observed that individuals moving into gentrifying neighbourhoods utilized social media to bolster their neighbourhood's identity status, a trend that intensified as gentrification advanced. Social media activity serves a crucial commercial function by showcasing and promoting shifting consumer demands and consumption patterns within gentrifying neighbourhoods, thus contributing to the co-creation of urban landscapes (Chang and Spierings 2023) . Furthermore, it serves as a pivotal means of reinforcing emerging gentrification processes and solidifying gentrified areas, particularly among gentrifiers (Friesenecker and Lagendijk 2021). Recent research has underscored the effectiveness of data-driven methodologies in identifying emerging gentrification trends (Gray, Buckner, and Comber 2023).

3. Enhancing Hedonic House Price Models (HHPMs) with neighbourhood and place characteristics. Historically, the classical conception within HHPMs has predominantly focused on attributing house value to physical property and location characteristics (Rosen 1974; Holmes et al. 2017). Property attributes encompass elements such as age, number of bedrooms, floor area, among others (Follain and Jimenez 1985). Meanwhile, location attributes typically revolve around access to amenities (schools, parks, shops, etc.) and distances to transportation hubs, workplaces, central business districts, etc (Osland 2010; Poudyal, Hodges, and Merrett 2009). In certain instances, they may also encompass variables like crime rates, ethnic composition, or pollution levels (Lynch and Rasmussen 2001; Hui et al. 2007). However, HHPMs have traditionally omitted measures related to "place". In this context, "place" refers not only to individuals' lived experiences (Agnew 2011) but also to the spaces where social relations occur. According to the classic conceptualization of place, these spaces are manifold, dynamic, and in a perpetual state of evolution (Massey 2002). Thus, while HHPMs have incorporated spatial characteristics of location, they have overlooked place-based attributes that encapsulate the dynamic and evolving nature of social relations. The integration of Spatial Media Data (SMD) into spatial and temporal analyses of hedonic house price data presents an opportunity to capture the evolving dynamics of place-making by incoming gentrifiers and the displacement experienced by those being marginalized.

Methods

A case study covering 3,712 km^2 in the north of England was selected to undertake the exploratory work. SMD was extracted from the Twitter API in batches of 5000, for the area over the period 2014-2019. Some 2.7 million geo-located tweets were downloaded, and their annual counts and distribution across 2049 LSOA neighbourhoods are shown in Table 1. LSOAs are Lower Super Output Areas, a standardised census reporting area with about 1500 residents. The LSOAs are here used to provide proxies for neighbourhood areas and to provide an indication of the relative frequency tweets. They are not used further in the analysis. The study area and the LSOAs are shown in Figure 1.

In many instances social media data are simply analysed using text mining, sentiment analysis and lexicons of terms to generate some measure of happiness or satisfaction. One of the aims of this research was to explore text engineering and natural language processing (NLP) to advance current approaches to SMD perception

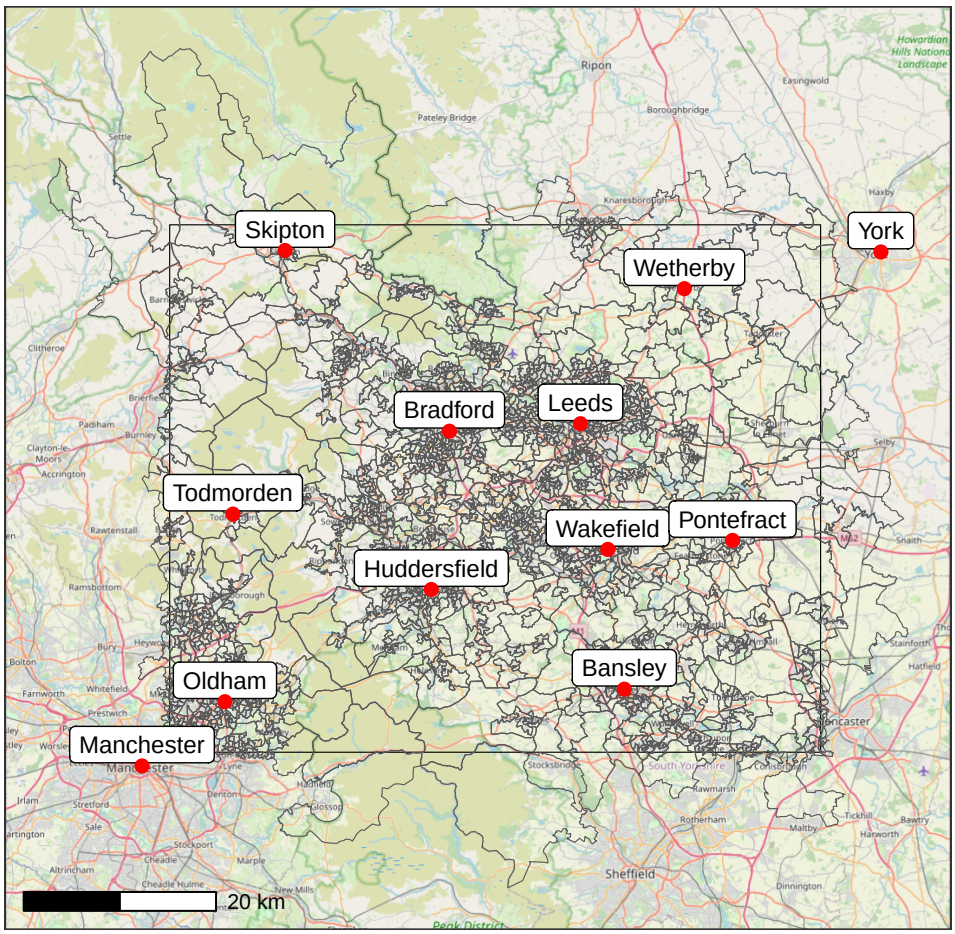

Figure 1: The study area with the 2049 LSOAs, some local towns and an OpenStreetMap backdrop.

analysis. An iterative approach was taken to explore ontologies and corpuses of concepts associated with gentrification-related neighbourhood change and their manifestation in SMD, and to quantify the relative advantages of NLP-grounded approaches over the more usual text mining ones. An initial analysis created a gentrification lexicon. This was done by copying whole blocks of text from descriptions, Wikipedia and online articles about gentrification. No text or words were extracted from academic articles describing gentrification. The aim here was explore how a sentiment-like analysis could be used, but with gentrification related words rather than words related to positive or negative sentiment. A random sample of the words associated with gentrification are shown in Table 2. The gentrification analysis simply scored each tweet with the number of times a word in the gentrification lexicon occurred in the tweet. For comparison, a sentiment analysis of positive and negative words was also undertaken using the lexicons from Prof Bing Liu's resources at University of Illinois at Chicago (http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar).

The analysis used the counts of gentrification-related words to construct and interpolated surfaces over a 500m grid. Here a standard inverse distance weighting approach was used and the gentrification scores for each year were interpolated over a 2 km^2, with a distance weighting power of 2, for to examine the sentiment scores for all years combined and a power of 4 for individual years to emphasise the spatial trends.

Results

The first set of results compares the gentrification score with standard sentiment scores. The results are shown in Table 3. Generally there is moderate positive correlation (~0.3) between the gentrification scores and positive sentiment scores and weak positive correlation with negative sentiment scores (~0.1). This

Table 2: A random sample of the 1,247 words associated with gentrification, used in the gentrification sentiment analysis.

16th	characteristic	discourse	gentrification	looking	precise	slum
1980s	comfort	disease	golden	lulu	preeminence	sprawl
according	concentrated	district	guiltridden	met	progressive	take
addition	continue	driven	heights	mortgages	pundits	thing
anti	contrast	due	historians	music	pursued	two
architecture	creatives	empirical	hot	needs	read	undersupply
array	creeping	enforcement	households	never	rent	unslumming
aspects	crime	even	imagined	noun	river	vehemently
attractions	dangerous	example	implicated	number	rome	venues
ballet	defines	federal	interpretation	paris	rose	whatever
believed	deprived	gains	lees	philosophical	rubric	wide
book	diffused	genderrelated	literal	place	seem	young

Table 3: The Correlations between sentiment and gentrifciation scores, 2014 to 2019.

Year	Postive-Gentrification	Negative-Gentrification
2014	0.301	0.187
2015	0.295	0.169
2016	0.274	0.039
2017	0.284	-0.057
2018	0.460	0.050
2019	0.480	0.098

indicates that the gentrification score is saying something different to simple sentiment.

The interpolated SMD derived gentrification scores for all years (2014-2019) are mapped in Figure 2. This shows some distinct trends, that are not just a function of of population. For example discrete areas of high gentrification sentiment are found to the north west of Bradford, south west of Wakefield and Huddersfield and west of Barnsley. When the annual scores are examined, the pattern is less obvious and indicative of the need for methodological refinement. Figure 3 shows the rescaled annual gentrification sentiment scores for each year. The rescaled data for each year has a mean of 0 and and a standard deviation of 1 indicating the the relative within-map sentiment. Some of the spatial trends evident in Figure 1 can be seen but it is difficult to determine any temporal trends, although potentially some emergent gentrification is evident to the north east of the case study area between 2015 and 2017, tailing off by 2018, and to the north and north west from 2017.

Discussion

The aim of this paper was to explore the extent to which social media data from Twitter could be used to capture perceptions of neighbourhood characteristics. Typically research employing SMD uses some kind of sentiment analysis and examines the frequency of positive and negative words, to summarise sentiment over social media post, location or object of analysis. Here a rudimentary lexicon of words associated with gentrification was created, extracted from blogs, dictionaries and online articles, and used to give a gentrification score to geo-located tweets. These were then interpolated to create the surfaces in Figures 2 and 3 and some discrete spatial and temporal patterns identified.

There are a number of limitations to the methods used in this study. The first relates to the reliability of the gentrification scores, given the looseness of the sources. Future work will identify more robust definitions, terminology, phrasing from academic articles describing gentrification and from experts in this domain.

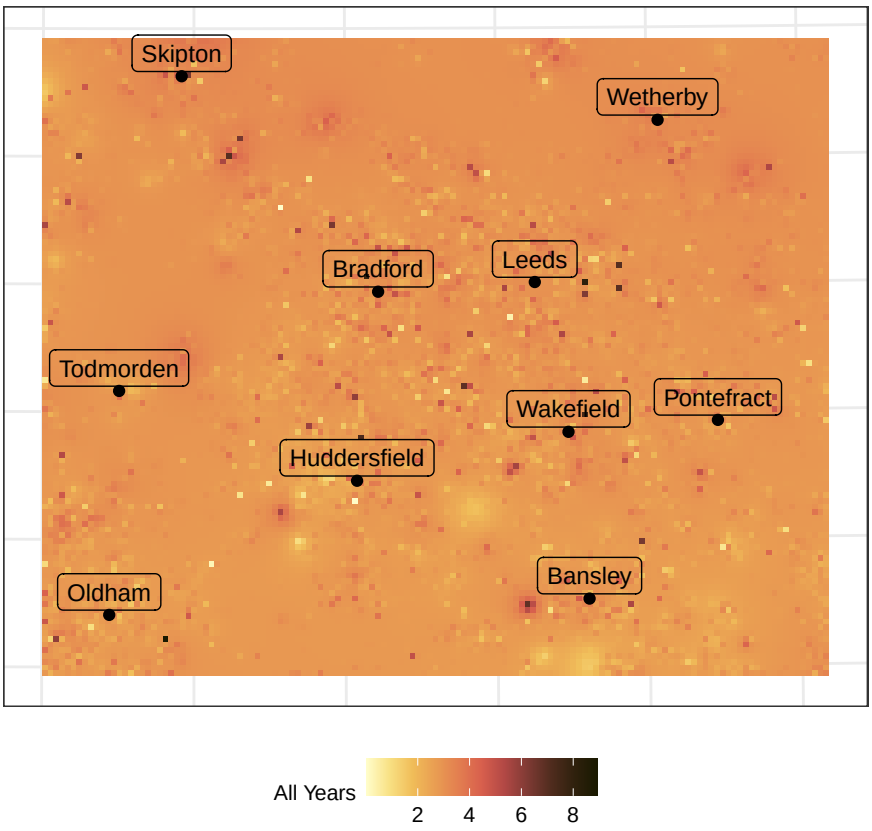

Figure 2: The overall interpolated gentrification scores, over all time years.

The second is the analysis of individual words, taken out of context and used to generate scores. This is problematic and a critical deficiency of all analyses of social media data employing text mining for anything other than classification. Future work will explore a refined natural language programming approach using the GATE (General Architecture for Text Engineering) interface hosted at the University of Sheffield (https://gate.ac.uk), to allow phrases and terms in context to be analysed as indicators of gentrification. Scores derived from these will in turn be linked to spatio-temporal house price data to construct predictive models of gentrification, neighbourhood change and house price bubbles.

The context for this work is that it is the first analysis step in a wider project that will link SMD-derived perceived neighbourhood characteristics from NLP to spatio-temporal house price data containing classic hedonic variables (e.g. property size, age, location, number of bedrooms etc.). The aim is to enhance HHPMs and to generate wider understanding of the influence of qualitative perceptions on house prices, their spatio-temporal dynamics in order to model neighbourhood changes. It will also examine the use of spatially and temporally varying coefficient analysis (Comber et al. submitted) to identify nascent house price bubbles, linked to increases in perceived neighbourhood quality, and thereby locales of emergent gentrification. While the analysis presented in this short paper focuses on West Yorkshire, UK, the full project will undertake parallel analyses in 3 other global, contrasting case studies in New Zealand and Vietnam. This will allow the use of SMD analysed in this way to be evaluated in different contexts. It will explore variations in representativeness and bias in SMD, the extent to which the perceptions of neighbourhood characteristics act as a driver of house price dynamics, how this varies internationally, with different social media platforms and in different linguistic contexts. The results of these activities will develop a deeper understanding of the robust and repeatable of use of SMD, and will extend data driven methods linked to explicitly spatial analyses to quantify neighbourhood dynamics.

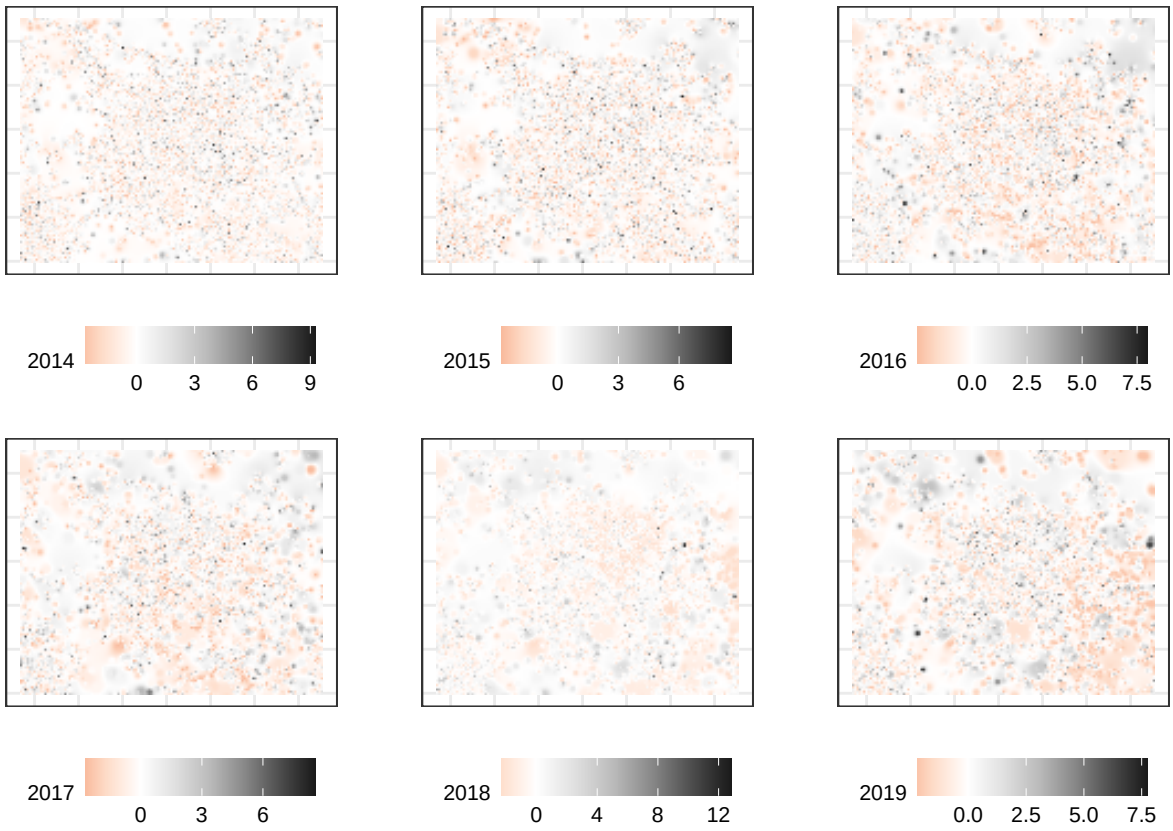

Figure 3: The rescaled interpolated gentrification scores, over each year 2014-2019.

Acknowledgements

This research is supported by UKRI (ESRC) funding ES/Y006259/1 under the Digital Footprints scheme. The code and some sample data have been made available at https://figshare.com/s/92c0a432d8087ec879e7, but it is not possible to share the full Twitter dataset due to its size and maybe licensing.

References

Agnew, John. 2011. "Space and Place." *Handbook of Geographical Knowledge* 2011: 316–31.

Bronsvoort, Irene, and Justus L Uitermark. 2022. "Seeing the Street Through Instagram. Digital Platforms and the Amplification of Gentrification." *Urban Studies* 59 (14): 2857–74.

Chang, Hanbit, and Bas Spierings. 2023. "Places 'for the Gram': Millennials, Specialty Coffee Bars and the Gentrification of Commercial Streets in Seoul." *Geoforum* 139: 103677.

Comber, Alexis, Paul Harris, Naru Tsutsumida, Jennie Gray, and Chris Brunsdon. submitted. "Where, When and How? Specifying Spatially and Temporally Varying Coefficient Models Using GAMs with Gaussian Process Splines." *International Journal of Geographical Information Science*, submitted.

Follain, James R, and Emmanuel Jimenez. 1985. "Estimating the Demand for Housing Characteristics: A Survey and Critique." *Regional Science and Urban Economics* 15 (1): 77–107.

Friesenecker, Michael, and Arnoud Lagendijk. 2021. "Commercial Gentrification in Arnhem and Vienna: The Local Enactment of, and Resistance Towards 'Globally'circulating Creative Neighbourhood Policies." *City* 25 (5-6): 698–719.

Gibbons, Joseph, Atsushi Nara, and Bruce Appleyard. 2018. "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): 470–88.

Gray, Jennie, Lisa Buckner, and Alexis Comber. 2021. "Extending Geodemographics Using Data Primitives: A Review and a Methodological Proposal." *ISPRS International Journal of Geo-Information* 10 (6): 386.

———. 2023. "Identifying Neighbourhood Change Using a Data Primitive Approach: The Example of Gentrification." *Applied Spatial Analysis and Policy* 16 (2): 897–921.

Holmes, Thomas P, Wiktor L Adamowicz, Fredrik Carlsson, Patricia A Champ, Kevin J Boyle, and Thomas C Brown. 2017. "A Primer on Nonmarket Valuation: The Economics of Nonmarket Goods and Resources."

Hui, Eddie CM, Chi Kwan Chau, Lilian Pun, and MY Law. 2007. "Measuring the Neighboring and Environmental Effects on Residential Property Value: Using Spatial Weighting Matrix." *Building and Environment* 42 (6): 2333–43.

Lansley, Guy, and Paul A Longley. 2016. "The Geography of Twitter Topics in London." *Computers, Environment and Urban Systems* 58: 85–96.

Lynch, Allen K, and David W Rasmussen. 2001. "Measuring the Impact of Crime on House Prices." *Applied Economics* 33 (15): 1981–89.

Massey, Doreen. 2002. "Living in Wythenshawe." In *Unknown City: Contesting Architecture and Social Space*, edited by Pivaro A Borden I Kerr J, 459–75. MIT Press Cambridge, MA.

Nguyen, Quynh C, Suraj Kath, Hsien-Wen Meng, Dapeng Li, Ken R Smith, James A VanDerslice, Ming Wen, and Feifei Li. 2016. "Leveraging Geotagged Twitter Data to Examine Neighborhood Happiness, Diet, and Physical Activity." *Applied Geography* 73: 77–88.

Osland, Liv. 2010. "An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling." *Journal of Real Estate Research* 32 (3): 289–320.

Poorthuis, Ate, Taylor Shelton, and Matthew Zook. 2022. "Changing Neighborhoods, Shifting Connections: Mapping Relational Geographies of Gentrification Using Social Media Data." *Urban Geography* 43 (7): 960–83.

Poudyal, Neelam C, Donald G Hodges, and Christopher D Merrett. 2009. "A Hedonic Analysis of the Demand for and Benefits of Urban Recreation Parks." *Land Use Policy* 26 (4): 975–83.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82 (1): 34–55.

Walters, Peter, and Naomi Smith. 2024. "It's so Ridiculously Soulless: Geolocative Media, Place and Third Wave Gentrification." *Space and Culture* 27 (1): 94–109.