Characterising neighbourhood dynamics through social media anlaysis and house sales transactions

A Comber[1]*, M Asher[1], Y Wang[1], M Kieu[2], BT Quang[3], HNT Thuy[4], PH Huu[5], N Malleson[1,6]

[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, Hanoi, Vietnam
[4]VNU Vietnam Japan University, Hanoi, Vietnam
[5]R&D Consultants, Hanoi, Vietnam
[6]Leeds Institute for Data Analytics, University of Leeds, UK
*email: a.comber@leeds.ac.uk

Abstract

This paper describes a two stage approach for identify neighbourhood areas that may undergoing gentrification related changes. It summarises classic hedonic house price data over time (2014-2023) for each neighbourhood, and compares neighbourhood average price with those of local nearby areas. This enables neighbourhoods experience high relative increases in price to be identified as potentially gentrifying areas. Social media data for these areas were extracted and analysed using a large language model and used to score individual social media posts with the a measure of the degree to which their content indicates that the neighbourhood is experiencing change, providing confirmatory evidence or not of gentrification. A number of areas of further work are identified.

Key Words: neighbourhood dynamics, house price model, Twitter

This paper will be formatted to the submission template if accepted for oral presentation - we just ran out of time!

Introduction

This paper describes ongoing work under the Integrated Analysis of Social Media and Hedonic House Prices for Neighbourhood Change (INTEGRATE) project (https://urban-analytics.github.io/INTEGRATE/intro.html). The high level project aim is to determine whether useful and reliable information about neighbourhood dynamics can be extracted from social media data (SMD), and the project conceit is to link the results of SMD analyses to hedonic house price (HHP) data in order to identify neighbourhoods whose character and composition is changing, and that are potentially subject to gentrification processes. The rationale for exploring SMD in this way is to determine the degree to which usable information about population preferences and behaviours can be extracted from SMD and used to augment other statistical models. Classic HHP models classically consider only quantifiable property characteristics (age, bedrooms, area), area related ones such as amenities, schools, parks and shops, pollution, crime demographic composition and utility consideration such distance to work or transport hubs (Follain and Jimenez 1985; Osland 2010; Poudyal, Hodges, and Merrett 2009; Lynch and Rasmussen 2001; Hui et al. 2007).

However, these fail to capture people's lived experience and sense of place (Agnew 2011; Massey 2002) as well as what Huu Phe and Wakely (2000) to the *intangible* characteristics of neighbourhoods. These combine to create different status poles - locales with different levels of attractiveness to different demographic groups. These are formed through the intersection of property and neighbourhood physical characteristics and resident perceptions perceptions (Comber et al. 2016) through what Huu Phe and Wakely (2000)

referred to as *Status-Quality-Trade-Off* (SQTO). Analyses of SQTO are traditionally supported by street level information, captured through extensive neighbourhood surveys. Although thematically detailed, these are temporally static and expensive.

The INTEGRATE project is exploring how SMD can be used to inform HHP models, using a range of different SMD sources, linked with HHP data and applied in different urban contexts (UK, New Zealand, Vietnam). Gentrification provides a convenient lens through which to do this, and this paper describes an analysis of neighbourhoods in the UK, with some detail around a local case study. This analysis is undertaken using a 2 stage process to identify neighbourhoods whose character and composition have changed over time, and that are potentially subject to gentrification processes. The first part identifies locales of potential neighbourhood change by comparing house price changes over time of individual neighbourhood areas with the changes in nearby areas. Then SMD (Twitter / X) is analysed using a large language model (LLM) to determine whether any confirmatory signals of gentrification change can be discerned.

Methods and Data

A two stage analysis was undertaken. First, an analysis of house price data over neighbourhood areas to identify areas where house prices had risen in much greater proportion to those in neighbouring areas. Second to extract coincident SMD from Twitter for those areas and then to analyse it using large language models.

Records of ~8.9 million house sales transactions (2014 to 2023) from WhenFresh/Zoopla and obtained from the CDRC (https://data.cdrc.ac.uk), was located by the property postcode. This in turn allowed each property sales to be located with neighbourhood areas. In this case Lower Super Output Areas (LSOAs) for England were used as the neighbourhoods ($n = 32,844$). These are the second finest resolution over which census data are reported (not used in this analysis) and each LSOA contains around ~1,500 people (~500 households). The annual median house price for each LSOA was calculated and then compared the annual local median house price calculated over the k-nearest neighbours approach, with $k = 250$. The 10 annual ratios were used to construct a regression against time for each LSOA, to determine the rate of change of local median house price to nearby median house price over time (β_{time}). The coefficient estimates calculated in this way were used to identify potential gentrifying neighbourhoods - that is, those experiencing much greater increases in house prices, relative to other neighbourhoods in their local area. An example of the nearest neighbours is shown in Figure 1 along with the distribution of the coefficient estimates. The latter indicates a normal distribution confirming the suitability of using standard deviations (here $> 2\sigma$) to identify potentially gentrifying neighbourhoods. The development of this method for identifying potentially gentrifying, is described elsewhere (Comber et al. 2025) using Medium Super Output Areas (MSOAs) which contain around 7,500 people (~2,500 households). Here this method was applied to examine LSOAs and to target the analysis of social media data, that has the potential to provide confirmatory evidence of gentrification.

Having identified potentially gentrifying areas, social media data was examined for indications of gentrification. Some 18 million tweets were downloaded from the Twitter / X API for the Leeds and Bradford case study area for the period 2014 to 2019 (i.e. prior to the COVID-19 pandemic). Social media posts (Tweets) for the potentially gentrifying areas were extracted and then passed to a LLM to investigate the degree to which gentrifying sentiment could be identified in these areas. Each of the tweets was analysed using the Meta Llama 3.3 70B Instruct Turbo LLM, accessed using an API provided by the service together.ai (https://www.together.ai). The following illustrates the system prompt that is passed to the LLM, developed with help from ChatGPT, as described in related work reported elsewhere (Malleson et al. 2025). The LLM was prompted with the following text:

> You have a deep understanding of neighbourhood character and how it is experienced discussed in public discourse. I will provide you with some Twitter / X posts ("tweets"). Your task is to analyse each one text and determine the extent to which each Tweet suggests that the neighbourhood is experiencing change. Specifically:
>
> Read the posts closely and identify any words, phrases, or implications that might indicate signs of neighbourhood change, changing demographics, or neighbourhood 'revitalisation'.

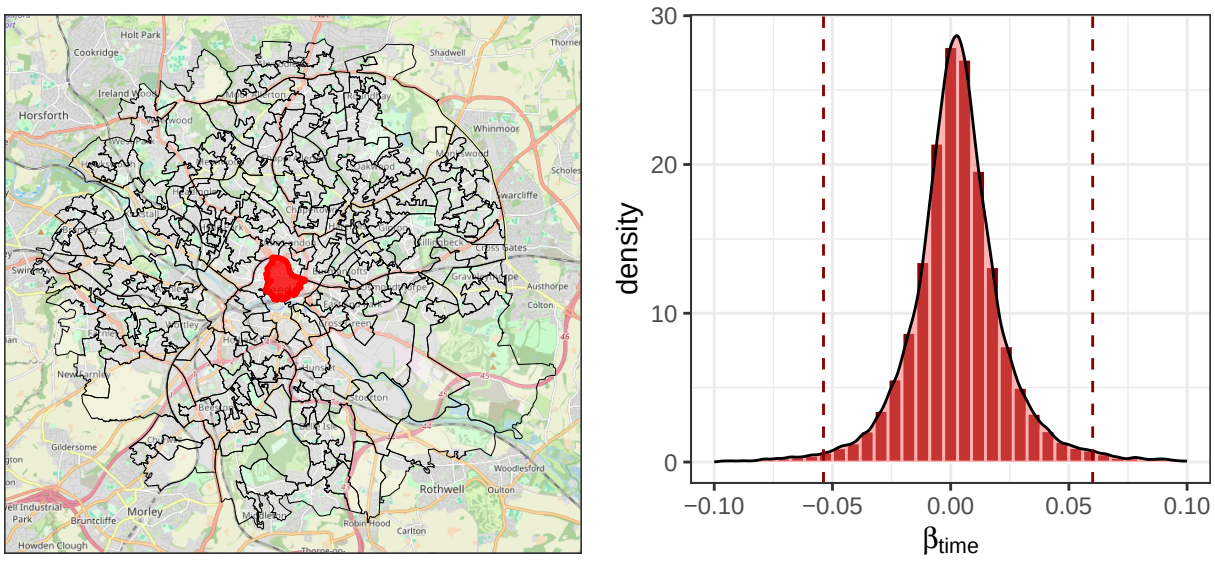

Figure 1: An example of an LSOA (500 households) in Leeds, UK and its 250 nearest neighbouring LSOA areas, with an OpenStreetMap backdrop (left), and A density histogram of the distribution of the coefficient estimates of the rate of change the ratio of LSOA house price to local nearby house prices over time 10 year period (2014 to 2023), with 2 standard deviations from the mean indicated (right).

> Consider both explicit and implicit cues. Explicit cues directly mention new businesses or rising prices, while implicit cues might reflect subtle neighbourhood changes.
>
> Assign a score from 1 to 5, where 1 means not suggestive of change and 5 means highly suggestive.
>
> Do not explain any reasoning.
>
> Provide your answer strictly in the format '1. Score', '2. Score', '3. Score', etc., without any additional explanation or commentary.

Results

The firs part of the analysis sought to identify potentially gentrifying areas from house price changes. These were LSOAs whose house price trajectories (regression slopes) over time, when quantified as a ratio to those of their 250 nearest neighbour areas, were greater than 2 standard deviations above the national mean. A total of 10 potentially gentrifying areas in the Leeds and Bradford case study area were identified. These are highlighted in Figure 2. These are candidate gentrifying areas.

In the second part of the analysis, spatially coincident Twitter data (2014 to 2019) for each of these areas were extracted and then passed to a Large Language Model (LLM). Here the task was to investigate the degree to which gentrifying sentiment could be identified in these areas. Some 92,411 tweets were coincident for the 10 areas over the 6 years but with very different distributions over time and space (Table 1).

Table 1: Counts of the tweets collected over the 10 potentially gentrifying LSOA areas, 2014 to 2019.

LSOA	2014	2015	2016	2017	2018	2019
E01010579	3,422	5,000	5,000	5,000	5,000	5,000
E01010693	1,865	655	59	84	48	16
E01010696	4,268	712	145	122	111	73
E01011270	5,000	4,579	2,773	2,984	4,373	5,000
E01011324	5,000	3,014	671	387	285	324

LSOA	2014	2015	2016	2017	2018	2019
E01011457	3,206	796	227	87	79	63
E01011561	5,000	1,878	0	0	0	0
E01011647	3,774	429	153	30	21	16
E01011668	4,264	1,270	88	27	21	12

These were passed to a LLM and scored for sentiment that suggested that the neighbourhood was experiencing change. The results are summarised in Table 2. Of particular interest are LSOAs with the codes `E01011324`, `E01011647` and `E01011668` (highlighted). These are mapped in their local context Figure 3.

Table 2: The mean neighbourhood change scores of the tweets collected over the 10 potentially gentrifying LSOA areas, 2014 to 2019.

LSOA	2014	2015	2016	2017	2018	2019
E01010579	1.13	1.11	1.19	1.24	1.21	1.20
E01010693	1.13	1.18	1.53	1.02	1.31	1.06
E01010696	1.26	1.17	1.46	1.40	1.96	1.42
E01011270	1.13	1.15	1.25	1.29	1.17	1.08
E01011324	**1.18**	**1.27**	**1.67**	**1.95**	**1.66**	**1.85**
E01011457	1.12	1.31	1.31	1.15	1.29	1.43
E01011561	1.14	1.24	NA	NA	NA	NA
E01011647	**1.14**	**1.28**	**1.41**	**1.80**	**1.95**	**1.88**
E01011668	**1.12**	**1.14**	**1.39**	**1.22**	**1.90**	**1.67**

Discussion

This paper presents 2 stage approach for linking information from house sales data and social media data. Initial work (reported in Comber et al. (2024)) took the approach of mining SMD to identify neighbourhoods that were potential subject to gentrification related changes. This used a simple text analysis of SMD over time, which scores social media posts against sentiment and a basic gentrification lexicon. However, as the SMD was not collected using tags or keywords, and captures whatever social media users post, it is highly non-specific in nature such that any potential signal of gentrification was lost in the noise. A different, more targeted approach was needed. Here, analysis of house price data over time was used to identify potentially gentrifying areas. These were neighbourhoods that experienced high house price increases (greater than 2 standard deviations) relative to surrounding areas, using a k-nearest neighbour approach, with $k = 250$. This determination of k probably and this approach to defining *nearby* needs a stronger justification, but was selected because it captured an intuitive range that households would plausibly relocate within, capturing a population of about 375,000. Similarly, the choice of 2 standard deviations as a threshold was somewhat arbitrary, and future work will explore the degree to which an approach using continuous house price increase values could be applied as well as the associated additional analysis overheads. However, we are confident that this two stage approach can be effective in identifying neighbourhoods experiencing change, including gentrification.

The work above has identified areas that are "on the up": positive changes in house prices, and expression of neighbourhood change extracted form social media data. It also possible to extract areas experience the inverse as shown in Figure 4, areas that are potentially experiencing decline for different reasons. Future work will also explore these and the associated neighbourhood related sentiment being expressed in these areas.

Future work will also explore a number f other related aspects: Here median house price for each LSOA was used but different neighbourhoods often have very different types of housing stock (detached, terraced,

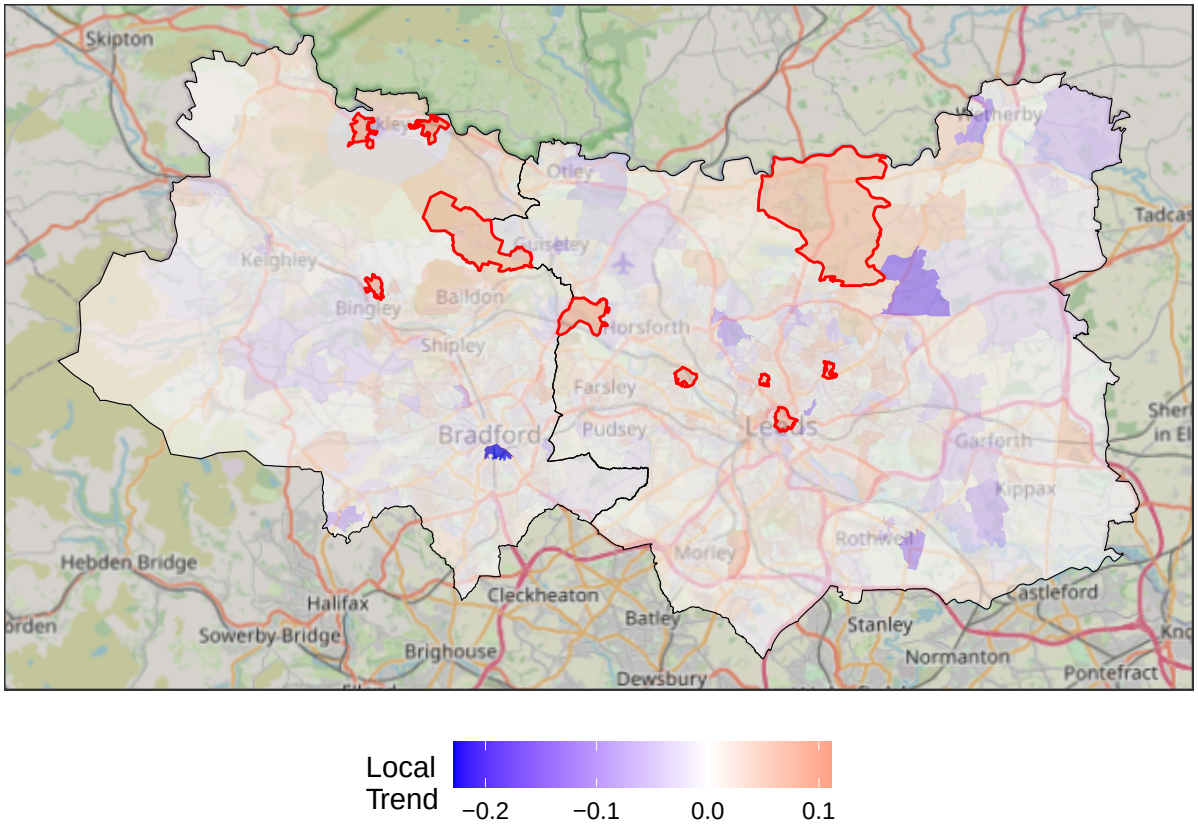

Figure 2: The coefficient estimates of local relative house price changes over time, compared to nearby areas (n = 250) for LSOAs in the Leeds and Bradford local govenrment areas (n = 792), with potentially gentrifying areas idnticated in red, and an OpenStreetMap backdrop.

Figure 3: The LSOAs with high relative house price increases with neighbourhood change social media scores gentrifying areas with an OpenStreetMap backdrop (left: E01011324, centre: E01011647, right: E01011668).

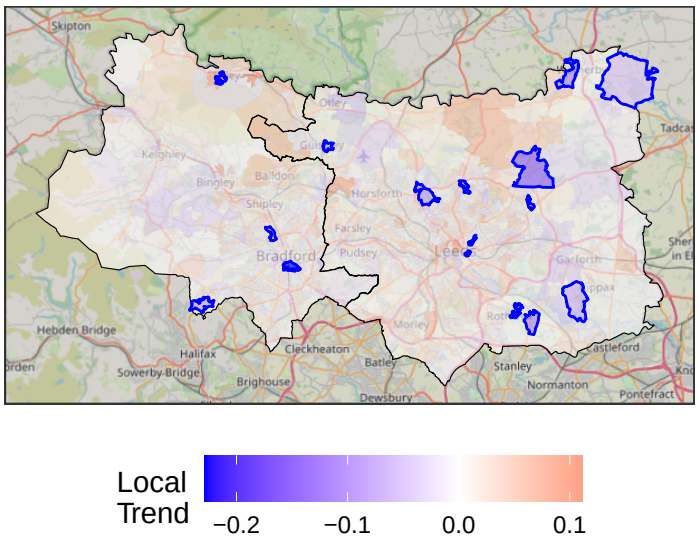

Figure 4: Relative house price changes over time for LSOAs as above but with the areas experiencing the greatest decreases in blue, and an OpenStreetMap backdrop.

apartments, etc), for which average prices can also be generated. Also *price per bedroom* may generate a more better measure with which to examine changes in house price.

> "Events, my dear boy, events", Harold Macmillan, 1963

Finally, when the INTEGRATE project was first proposed, we anticipated using standard techniques for examining social media data content for neighbourhood related sentiment using classic tools like text analysis, for example through the construction and application of a gentrification sentiment lexicon, and more advanced approaches like Natural Language Processing (NLP). These were specified in the proposal that was funded. However, events have overtaken us: in March 2023 the project team became aware of the potential of large language models and how quickly they could be prompted to generate sensible outputs (it took one of us 10 minutes to generate an acceptable undergraduate level essay), and subsequently their rapid improvements. In our work since then we have used LLMs for text classification, to generate and translate code, to interpret on-line reviews and now for social media analysis. They have become a standard part of everyday toolkits, in a way that could not have been anticipated when we wrote the proposal for this work. This has been the both an amazing opportunity and and a challenge for the project team.

Acknowledgements, DSA, Generative AI and Contributions

The authors contributed the ideas, theory and writing in this paper to differing degrees. AC did the bulk of the thinking, data collection and analysis. NM provided the interface into the LLM, with other authors contributing to discussions and brainstorming activities. No generative AI was used in the writing or ideas. This research is supported by UKRI (ESRC) funding ES/Y006259/1 under the Digital Footprints scheme. The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC, ES/L011840/1; ES/L011891/1. The data used in this analysis cannot be shared publicly, but this may be possible on request. Please contact the authors.

References

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

Comber, Alexis, Molly Asher, Yiyu Wang, Minh Le Kieu, Bui Thanh Quang, Hang Nguyen Thi Thuy, Phe Hoang Huu, and Nick Malleson. 2025. "Using House Sales Transactions Data to Identify Potentailly Genetrifying Neighbourhoods." In *Proceedings of 33rd Annual GIS Research UK Conference (GISRUK 2025)*.

Comber, Alexis, Paul Harris, Quan Nguyen, Khanh Chi, Hung Tran, and Hoang Huu Phe. 2016. "Local Variation in Hedonic House Pricing in Hanoi, Vietnam: A Spatial Analysis of Status Quality Trade-Off (SQTO) Theory." In *International Conference on GIScience Short Paper Proceedings*. Vol. 1. 1.

Comber, Alexis, Minh Kieu, Quang-Thanh Bui, and Nick Malleson. 2024. "Using Social Media Data to Identify Neighbourhood Change." *AGILE: GIScience Series* 5: 20.

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.

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.

Huu Phe, Hoang, and Patrick Wakely. 2000. "Status, Quality and the Other Trade-Off: Towards a New Theory of Urban Residential Location." *Urban Studies* 37 (1): 7–35.

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

Malleson, Nick, Bui Thanh Quang, Hang Nguyen Thi Thuy, Phe Hoang Huu, Minh Kieu, and Alexis Comber. 2025. "Using Large Language Models to Predict Neighbourhood Change." In *Proceedings of 33rd Annual GIS Research UK Conference (GISRUK 2025)*.

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.

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

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.