Leveraging recent advances in Large Language Models for neighbourhood delineation in London housing market

Vladimir R. Tesniere[*1], Stephen Law[†1]

[1]UCL, Department of Geography

GISRUK 2024

Summary

Neighbourhoods and their associated area are complex and spatial in nature, holding perceptual meaning for a city's inhabitants. However, when it comes to defining them, there is little consensus. To address this issue, we propose a methodology for constructing local housing area boundaries, which can enhance our spatial and temporal understanding of neighbourhoods in the housing-market. We leverage on recent advances in Large Language Models, specifically the BART model for Zero-shot classification and unsupervised Topic modelling using SBERT, to assign property listing data into to their respective housing market area. We validate our neighbourhood boundaries by comparing them to existing delineations derived from the Ordnance Survey Locality data.

KEYWORDS: Large Language Models, Zero-Shot Learning, Geo-text Analysis, Neighbourhoods, Housing-market

1.0 Background

The concept of a neighbourhood is complex, encompassing perceptual constructs and socio-economic characteristics (Law, 2017). For housing market, residential neighbourhoods play a critical role for real estate agents and home seekers to identify where to sell or buy a property. Currently, these housing neighbourhood boundaries lack consensual method and are prone to observer bias when constructed manually. Utilizing the increased availability of geo-text data (Hu, 2018) and recent advances on large language models (Brown et al 2020), the study proposed a novel zero-shot encoding pipeline that delineate housing marketing neighbourhood boundaries for London, UK drawing on data from OS and Houseful.

2.0 Datasets

This work leverages from three data sources:

1. H1 2023 sold property listings from the Houseful group, consisting of over *100k* observations, all paired with a unique reference (*UPRN*).
2. Web-scrape a collection of named areas in London with the potential of being a neighbourhood. These are collected from publicly available data-sources such as Wikipedia and OpenStreetMap, then using GeoHack® to provide coordinates for these locations. In total, we identify 532 potential areas in Greater London and use these as reference categories for classification in our model.
3. Ordnance Survey UPRN dataset, where some properties contain 'locality names'. From this we extrapolated ground truth Neighbourhood delineation for certain London Localities.

3.0 Methodology

This research applies the Bidirectional AutoRegressive Transformers or BART (Lewis et al 2019) pretrained with the Multi-Genre Natural Language Inference (MultiNLI) corpus as a zero-shot learner. The encoder-decoder model has strong semantic text recognition for multiple topics, making it a viable approach for location identification using geo-text data.

[*] stnvvrt@ucl.ac.uk
[†] stephen.law@ucl.ac.uk

Figure 1 illustrates the workflow of our model implementation. We begin by sub-setting both the listing data and the web-scraped neighbourhood names for each borough. Focusing on boroughs reduces classification errors when inputting listing data into the zero-shot encoding model (approximately 17 areas per property). For each property listing, we then estimate the neighbourhood with the highest probability using BART as a zero-shot learner. Only classifications above a score of 0.9 are kept. We consider the threshold at 90% as results rapidly deteriorate below such values (Barker et al., 2021). The classified outputs are merged with the property information which provide a full over-view of the property characteristics.

Once the listings had been classified, a housing market boundary is created by applying kernel density estimation (bandwidth=0.01) on the precise geo-location of every listing in our dataset. To prevent outliers in our final polygon creation, we considered the smallest area within the the Density estimate outline that captures at least 90% of the variance. KDE boundaries are then spatially joined with the Lower Layer Super Output Area (available here) that is within the proposed polygon as illustrated in **Figure 2**.

Figure 1: Workflow of the Zero-shot neighbourhood delineation model (left) and an example for property neighbourhood classification (right)

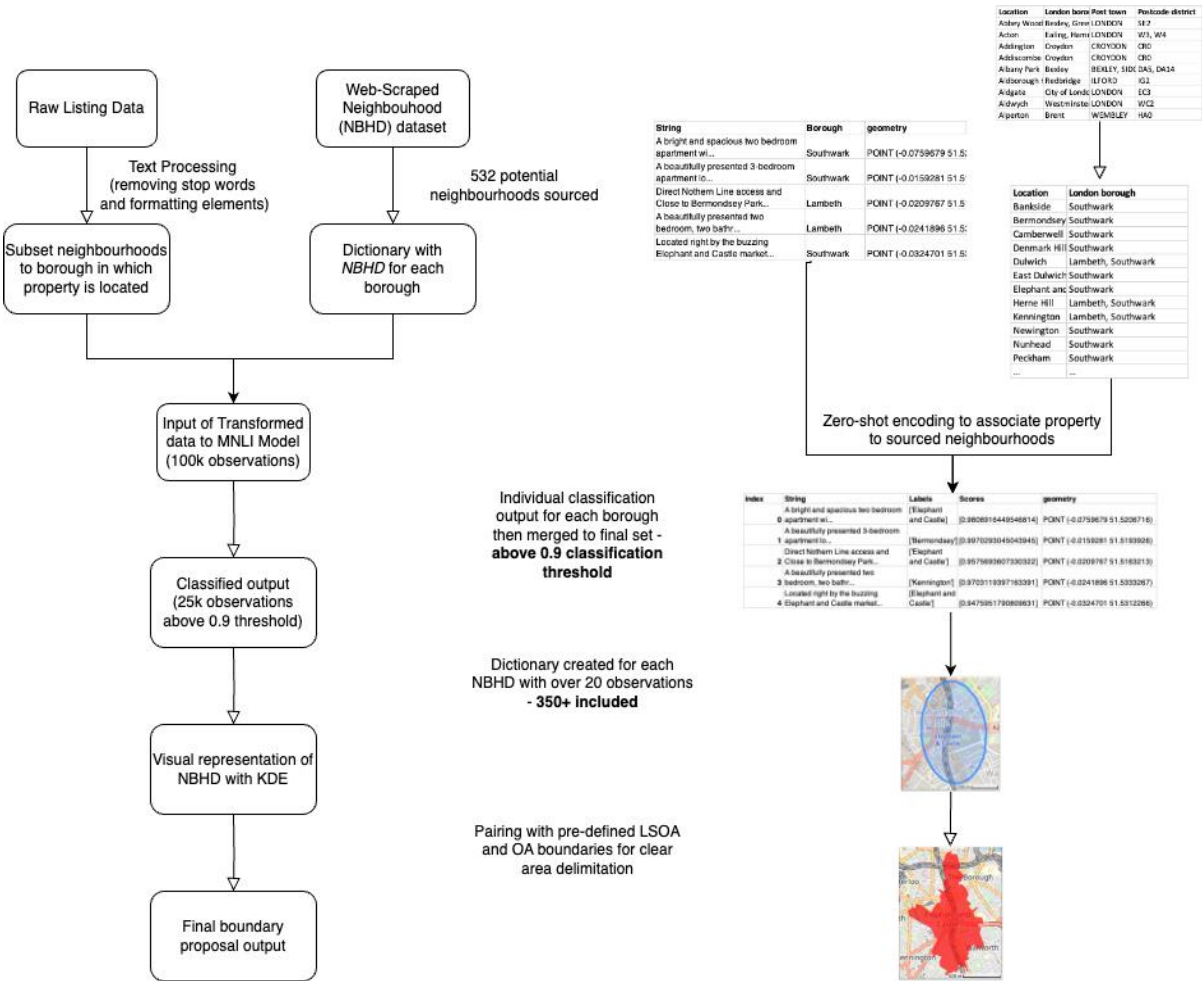
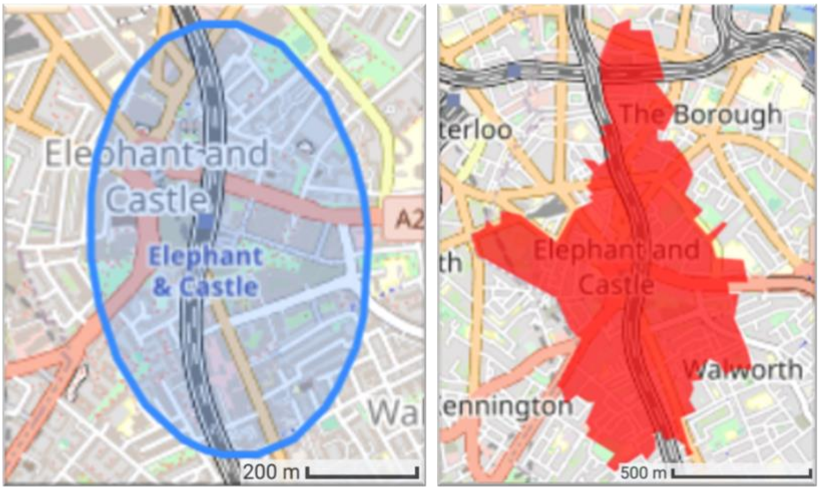


Figure 2: Initial model outputs (left) compared to combination with LSOA (right)



4.0 Results[‡]

Table 1: model overview, output, and total areas identified

Model	Unique listings	Classified listings (%)	Delineations proposed
Bart Zero-shot Encoding	102,844	28,576 (27.79%)	369 (of 532 total)
SBERT-DBSCAN Unsupervised clustering	102,844	15.001 (21.4%)	202
NER (spaCy)	102,844	8,231 (8.00%)	46

Table 1 indicates that the Zero-shot learner was able to identify 369 boundaries from the initial 532 neighbourhood names available in our area dataset. The areas with proposed delineations were the ones that met the classification requirements, where the zero-shot encoding score for the geo-text (listing) is above the 90% confidence threshold for more than 20 observations. It is estimated that for a dataset of low dimensionality, such as the one presented here, a minimum of 19 observations are necessary (Silverman, 1986). We compared this to a standard 'Named Entity Recognition' baseline using spaCy which only classified 8% of the listings, outputting 46 plausible areas.

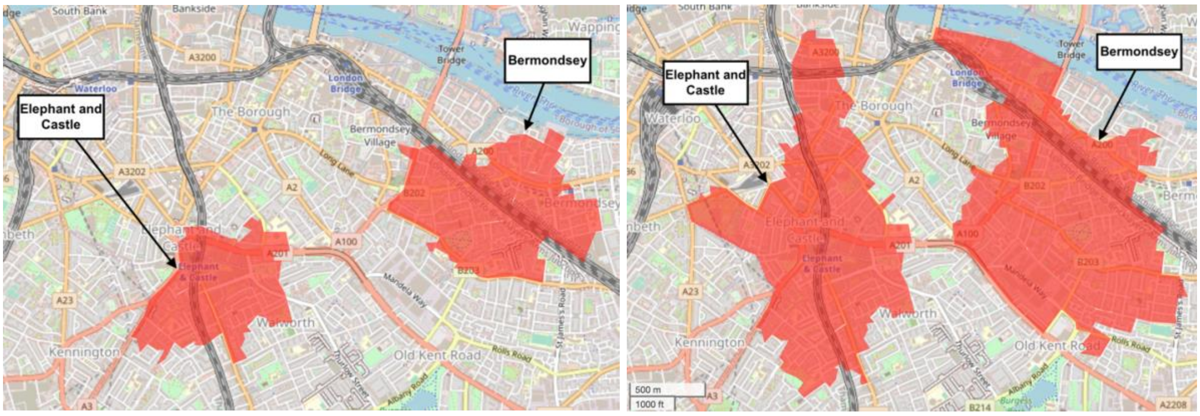
We also compared this with an unsupervised clustering approach which uses SBERT (Reimers and Gurevych 2019) to generate embeddings used for clustering with HDBSCAN (Schubert et al 2017). Once defined, TF-IDF is applied to the listings in each cluster to retrieve the most impactful tokens and bigrams. The results surpassed NER, with 21.4% of the listings classified and 202 neighbourhoods matching the impactful tokens. These results suggest that zero-shot encoding model was able to identify more neighbourhoods than an unsupervised clustering approach and NER method. An unsupervised approach can be useful when neighbourhood names are missing.

Figure 3 presents issues encountered in the testing phase with direct string match, aiming to extract location from explicit mentions in the listing. **Figure 4** illustrates additional results created using historical listing data, comparing neighbourhood evolution across time.

[‡] Additional findings and London-wide visualisations are available here

Figure 3: limitations in testing phase of direct string matching, through example of confused identification between Victoria Park and Victoria Station

Figure 4: illustration of neighbourhood evolution, delineation of Elephant & Castle and Bermondsey in 2015 (left) and 2023 (right)



4.1 Boundary Validation Methods

Given the absence of conventional validation metrics in neighbourhood delineation, we introduce a new metric to empirically validate our model outputs. We compare our boundaries with those provided by Ordnance Survey (OS) UPRN data, leveraging *"locality names"* in the dataset. Despite the OS dataset capturing only a limited number of neighbourhoods (150 for all greater London), it serves as a valuable ground-truth. Our validation focuses on the East-London area, using Intersection-over-Union (IoU) to compare boundaries. visually juxtaposes our model boundaries with OS-based polygons in East London. **Table 2** presents scores for area overlap and IoU. Using both metrics captures spatial similarities whilst also considering the importance of boundaries in a single measure (Maldonado and Zetzsche, 2023).

Figure 5: visual representation of model created boundaries and OS extrapolated

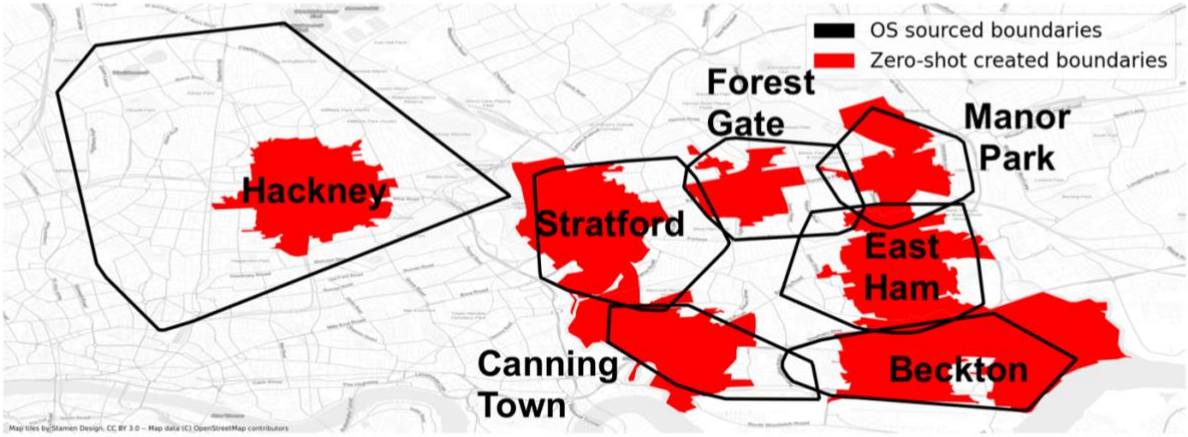


Table 2: comparing zero-shot defined boundaries to OS

Area Name (East London)	Overlap with OS source (%)	Intersection over union (IoU) between 0 and 1
Beckton	41.01	**0.57**
Canning town	68.72	**0.50**
Stratford	87.42	0.48
Manor Park	87.96	0.48
East Ham	96.71	**0.54**
Forest Gate	98.14	0.37
Hackney	100.0	0.15

An IoU score is considered significant when surpassing 0.5, in line with accuracy thresholds in image research domains (Yu *et al.*, 2018). Several areas have such scores, including East Ham, Canning Town, and Beckton. Conversely, Forest Gate exhibit higher overlap but lower IoU. Some areas such as Hackney display high overlap, as our proposed boundary is entirely with Ordnance Survey (OS) boundaries but a very low IoU implying Hackney serves as both a local neighbourhood and borough here. In general, the model is successful in identifying established areas, illustrating successful implementation for London which encourages further UK-wide classification.

5.0 Discussion

We propose a novel pipeline that uses Zero-Shot encoding to associate a property to a specific housing market area through analysis of listing description, *'geo-text'*, and then using this data to propose residential market area boundaries. The proposed method had been validated by comparing with two other methods and with ground truth boundaries.

The proposed framework captures spatial and temporal variations and is designed so that the model can continuously receive data as listings are created and recorded. The delineated areas created through Zero-shot classification exhibit overlap, which was expected due to the non-stationarity nature of neighbourhood boundaries. This would allow for precise visualisation of boundary evolution over time. Furthermore, these boundaries could provide 'gap-filling' for authoritative sources which lack granularity in urban area delineation.

Several limitations remain. LSOAs are advantageous for accounting for geo-demographic characteristics but they lack granularity therefore we propose implementing the boundaries from postcodes as they are the smallest area unit available in the UK. We acknowledge reproducibility concerns due to the 'BlackBox' nature of zero-shot encoding model, with potential classification changes following updates to the encoding model and training dataset. Intrinsic bias of the dataset is acknowledged as created listings aim to maximize property appraisals. This bias may spatially exaggerate popular areas and diminish less sought-after ones, creating geo-demographic disparities in information availability. We recognize the bias's impact on the model's potential recommendations to authoritative sources, but negligible in the housing market as it reflects current real estate trends.

In future works, we propose exploring rental market listings, characterized by higher fluctuation due to shorter tenures (Tomal and Helbich, 2022). The application of zero-shot encoding to rental listings is expected to capture differences in area sizes and the emergence or disappearance of areas.

Acknowledgements

We thank Zoopla and Houseful for their participation in this project and the ability to use their high quality and up-to-date data to fuel this research. We also thank the CDRC and UCL department of Geography for making this logistics of this project possible.

References

Barker, K., Awasthy, P., Ni, J. and Florian, R., (2021) NLI reranking for zero-shot text classification. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)* (pp. 193-202).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, pp.1877-1901.

Hu, Y. (2018) 'Geo-text data and data-driven geospatial semantics', *Geography Compass*, 12(11), p. e12404. Available at: https://doi.org/10.1111/gec3.12404.

Law, S. (2017) 'Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London', *Cities*, 60, pp. 166–179. Available at: https://doi.org/10.1016/j.cities.2016.08.008.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Maldonado, J. and Zetzsche, C. (2023) 'Representing (Dis)Similarities Between Prediction and Fixation Maps Using Intersection-over-Union Features', in *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. New York, NY, USA: Association for Computing Machinery (ETRA '23), pp. 1–8. Available at: https://doi.org/10.1145/3588015.3589843.

Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. Chapman & Hall, London. http://dx.doi.org/10.1007/978-1-4899-3324-9 (2018 edition)

Tomal, M. and Helbich, M. (2022) 'The private rental housing market before and during the COVID-19 pandemic: A submarket analysis in Cracow, Poland', *Environment and Planning B: Urban Analytics and City Science*, 49(6), pp. 1646–1662. Available at: https://doi.org/10.1177/23998083211062907.

Wessendorf, S. (2013) ‘Commonplace diversity and the “ethos of mixing”: perceptions of difference in a London neighbourhood’, *Identities*, 20(4), pp. 407–422. Available at: https://doi.org/10.1080/1070289X.2013.822374.

Yu, Z. *et al.* (2018) ‘Rethinking Diversified and Discriminative Proposal Generation for Visual Grounding’. arXiv. Available at: http://arxiv.org/abs/1805.03508 (Accessed: 9 September 2023).

Biographies

Dr. Stephen Law is currently a lecturer in Social and Geographic Data Science, researcher, and assistant professor at UCL and research fellow at the Alan Turing Institute. His research focuses principally on geographic data science, urban economic analysis, computer design and network science.

Vladimir R. Tesniere is a recent graduate from the MSc in Social and Geographic Data Science MSc at UCL. He has a background in spatial analysis, demographics, computing, and geographic data-science.