Leveraging advances in Large Language Models for neighbourhood delineation in the 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 BERTopic for unsupervised topic-modelling, to assign property listing data 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: LLM, Zero-Shot classification, 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) in GeoAI (Mai et al 2023), the study proposed a novel zero-shot classification pipeline for toponym recognition (Wang and Hu, 2019) in delineating 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

3.1 Zero-shot classification with BART

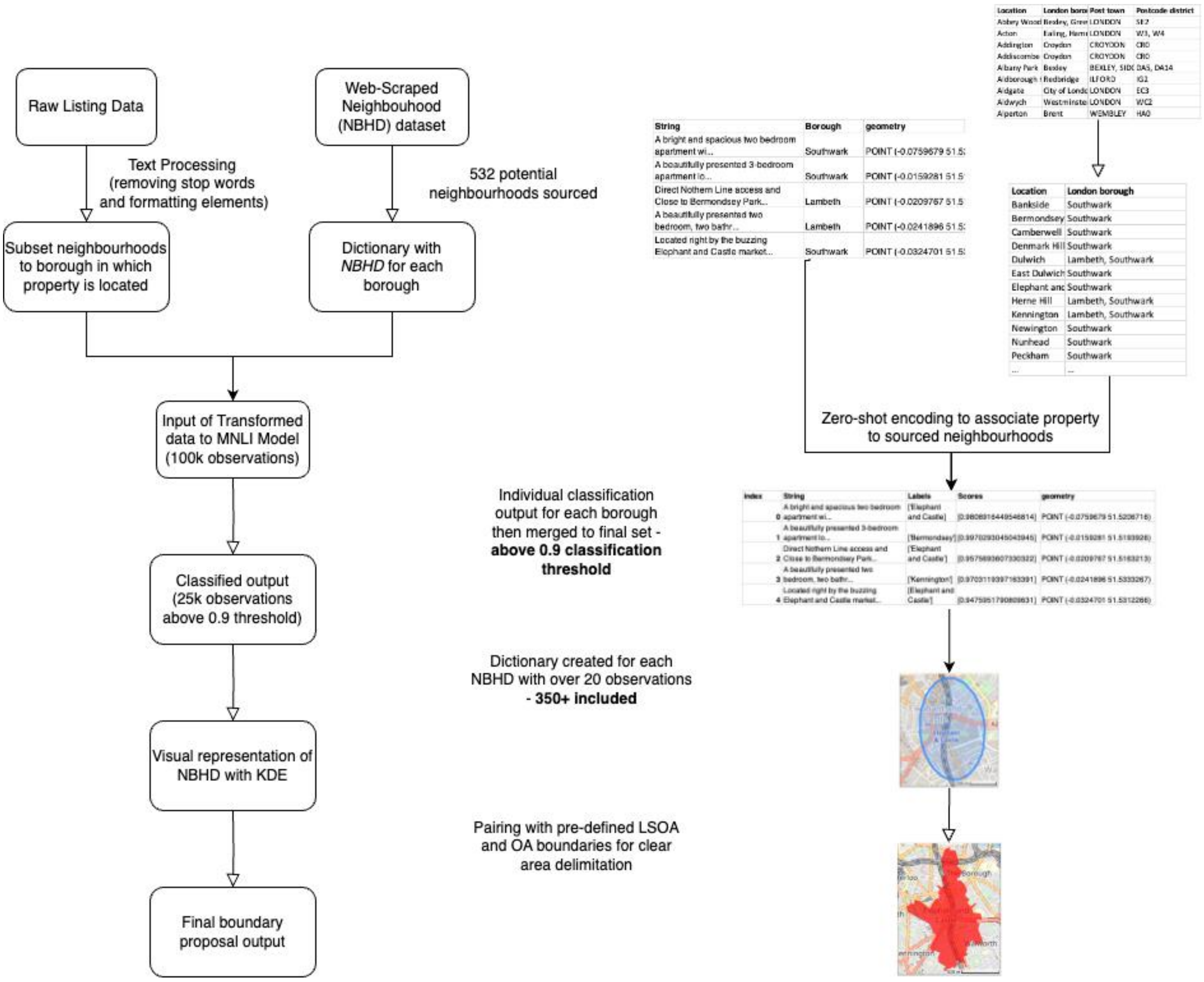
This research first applies the Bidirectional-AutoRegressive-Transformers or BART (Lewis et al 2019) pretrained with the Multi-Genre-Natural-Language-Inference (MultiNLI) corpus as a zero-shot learner. The encoder-decoder-transformer model has strong semantic text recognition for multiple topics, making 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 (*~17 areas/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 as results rapidly deteriorate below this threshold (Barker et al., 2021). The classified outputs are merged with the property information. 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)


Raw Listing Data
Web-Scraped Neighbouhood (NBHD) dataset
Text Processing (removing stop words and formatting elements)
532 potential neighbourhoods sourced
Subset neighbourhoods to borough in which property is located
Dictionary with NBHD for each borough
Input of Transformed data to MNLI Model (100k observations)
Classified output (25k observations above 0.9 threshold)
Visual representation of NBHD with KDE
Final boundary proposal output
Zero-shot encoding to associate property to sourced neighbourhoods
Individual classification output for each borough then merged to final set - above 0.9 classification threshold
Dictionary created for each NBHD with over 20 observations - 350+ included
Pairing with pre-defined LSOA and OA boundaries for clear area delimitation

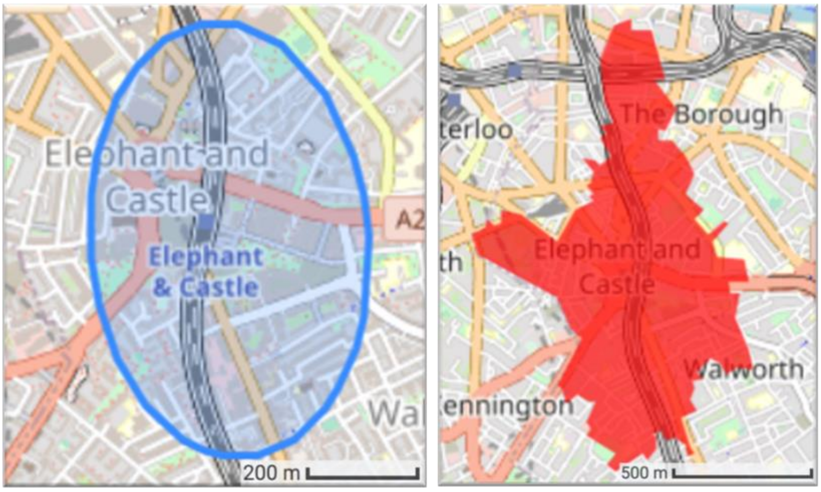
Location	London boro	Post town	Postcode district
Abbey Wood	Bexley, Gree	LONDON	SE2
Acton	Ealing, Hamr	LONDON	W3, W4
Addington	Croydon	CROYDON	CR0
Addiscombe	Croydon	CROYDON	CR0
Albany Park	Bexley	BEXLEY, SIDC	DA5, DA14
Aldborough (	Redbridge	ILFORD	IG2
Aldgate	City of Londc	LONDON	EC3
Aldwych	Westminste	LONDON	WC2
Alperton	Brent	WEMBLEY	HA0

String	Borough	geometry
A bright and spacious two bedroom apartment wi...	Southwark	POINT (-0.0759679 51.5:
A beautifully presented 3-bedroom apartment lo...	Southwark	POINT (-0.0159281 51.5
Direct Nothern Line access and Close to Bermondsey Park...	Lambeth	POINT (-0.0209767 51.5
A beautifully presented two bedroom, two bathr...	Lambeth	POINT (-0.0241896 51.5:
Located right by the buzzing Elephant and Castle market...	Southwark	POINT (-0.0324701 51.5:

Location	London borough
Bankside	Southwark
Bermondsey	Southwark
Camberwell	Southwark
Denmark Hill	Southwark
Dulwich	Lambeth, Southwark
East Dulwich	Southwark
Elephant anc	Southwark
Herne Hill	Lambeth, Southwark
Kennington	Lambeth, Southwark
Newington	Southwark
Nunhead	Southwark
Peckham	Southwark
...	...

index	String	Labels	Scores	geometry
0	A bright and spacious two bedroom apartment wi...	['Elephant and Castle']	[0.9808918449546814]	POINT (-0.0759679 51.5206716)
1	A beautifully presented 3-bedroom apartment lo...	['Bermondsey']	[0.9970293045043945]	POINT (-0.0159281 51.5193926)
2	Direct Nothern Line access and Close to Bermondsey Park...	['Elephant and Castle']	[0.9575693607330322]	POINT (-0.0209767 51.5163213)
3	A beautifully presented two bedroom, two bathr...	['Kennington']	[0.9703119397163391]	POINT (-0.0241896 51.5333267)
4	Located right by the buzzing Elephant and Castle market...	['Elephant and Castle']	[0.9475951790809631]	POINT (-0.0324701 51.5312266)

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



3.2 Unsupervised clustering with BERTopic

We also experimented with an unsupervised topic modelling technique. Specifically, a variant of BERTopic (Grootendorst 2022), which employs SBERT (Reimers and Gurevych 2019) for sentence embeddings and HDBSCAN (Schubert et al. 2017) for topic clustering. Once the clusters are defined, TF-IDF is then applied to the listings in each cluster to retrieve the most impactful tokens and bigrams in relation to named-neighbourhoods.

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-HDBSCAN Unsupervised clustering	70,000[§]	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 named-area dataset. The areas with proposed delineations were the ones that met the classification requirements. 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 tested the unsupervised SBERT-HDBSCAN, with 21.4% of the listings classified and 202 neighbourhoods matching the impactful tokens. These results suggest that zero-shot encoding model was the most performant follow by SBERT-HDBSCAN and NER (baseline) method.

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

[§] Unsupervised model used a subset of H1 2023 listings due to local computational limitations

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

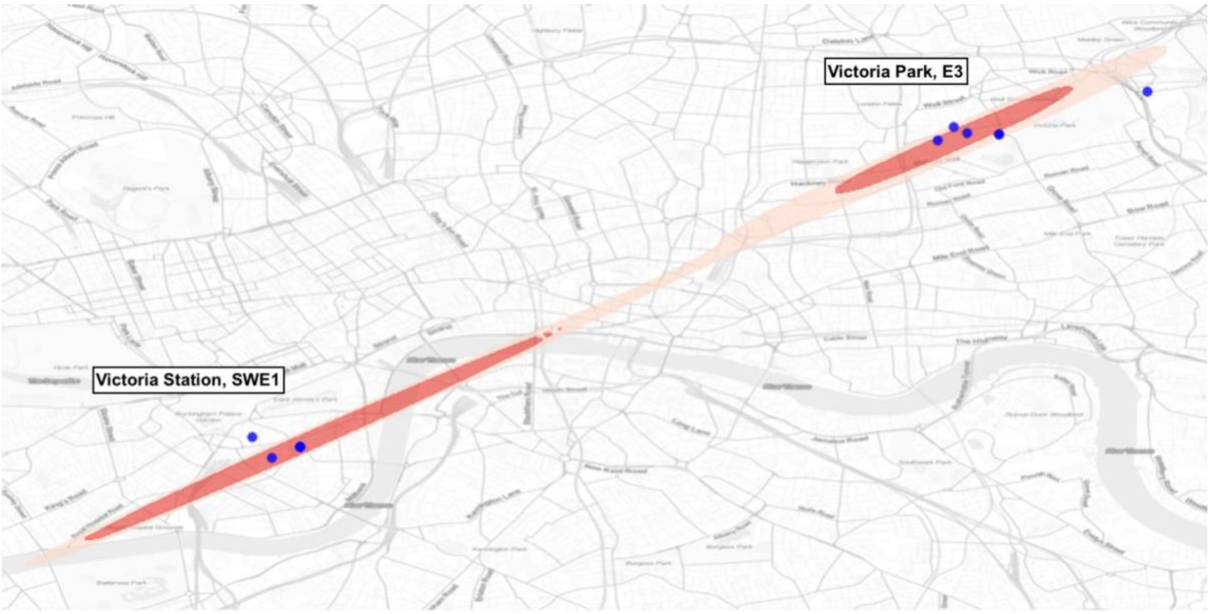
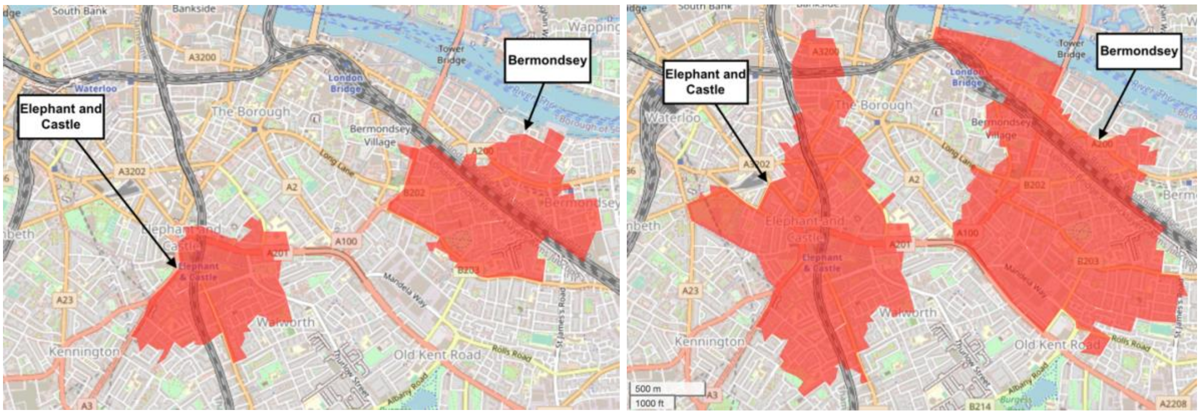


Figure 4: illustration of neighbourhood evolution, delineation of Elephant & Castle and Bermondsey in 2015 (left) and 2023 (right)



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. **Error! Reference source not found.** 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

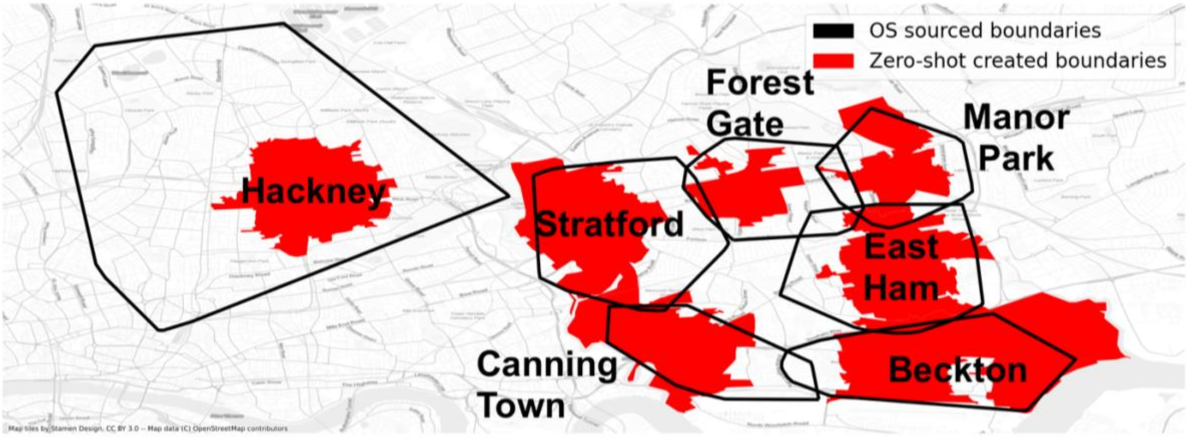

OS sourced boundaries
Zero-shot created boundaries
Hackney
Forest Gate
Manor Park
Stratford
East Ham
Canning Town
Beckton
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

Table 2: comparing zero-shot defined boundaries to OS ground-truths

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. Some areas such as Hackney display high overlap, as our proposed boundary is entirely within the 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 for London which encourages further UK-wide classification.

5.0 Discussion

We propose a novel pipeline that uses two large language modelling techniques namely, Zero-Shot-classification with BART and topic-modelling with BERTopic, 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 zero-shot-encoding method was more performant than the other methods which was then validated 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 exhibit overlap, which was expected due to the non-stationarity nature of neighbourhood boundaries. 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. We also aim to test capabilities of Zero-shot or Few-shot prompts for delineating housing market with popular LLM such as GPT4 (Mai et al 2023).

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 as part of the dissertation for the MSc Social and Geographic Data Science masters in UCL Geography. We also thank the CDRC and UCL Geography for making this project possible.

References

Barker,K.et.al.(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.et.al.(2020).Language models are few-shot learners.*Advances in neural information processing systems*,*33*,pp.1877-1901.

Grootendorst,M.(2022).BERTopic:Neural topic modeling with a class-based TF-IDF procedure.*arXiv preprint arXiv:2203.05794*.

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.et.al.(2019).Bart:Denoising sequence-to-sequence pre-training for natural language generation, translation,and comprehension.arXiv preprint arXiv:1910.13461.

Mai, G.et.al.(2023).On the opportunities and challenges of foundation models for geospatial artificial intelligence.*arXiv preprint arXiv:2304.06798*.

Maldonado, J.et.al.(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*.NewYork,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(2018edition)

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

Wang,J.,&Hu,Y.(2019).Enhancing spatial and textual analysis with EUPEG:An extensible and unified platform for evaluating geoparsers.*Transactions in GIS*,*23*(6),1393-1419.

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 preprint arXiv:1805.03508.

Biographies

Dr. Stephen Law is currently a lecturer (Assistant Professor) in Social and Geographic Data Science in UCL Geography. His research focuses principally on geographic data science, urban image analysis urban economics, urban design and space syntax.

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.