AGILE: GIScience Series, 7, 41, 2026. https://doi.org/10.5194/agile-giss-7-41-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

A graph-based community detection approach for identifying the semantic neighbourhoods within London's Airbnb properties

Joseph Shingleton [1], Yunus Serhat Bıçakçı [1,2], Yu Wang [1], Ana Basiri [1,3], and Meiliu Wu [1]

[1]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK
[2]Department of Artificial Intelligence and Machine Learning, Faculty of Applied Sciences, Marmara University, Istanbul, Türkiye
[3]Alan Turing Institute, London, UK

Correspondence: Joseph Shingleton (joseph.shingleton@glasgow.ac.uk)

Abstract. Neighbourhoods are a fundamental unit for organising, analysing, and understanding urban systems. While they can be described in administrative terms, subjective boundaries often better capture lived urban experience. The natural language people use to describe where they live provides one signal of these boundaries. We present a spatio-semantic approach to neighbourhood identification, recovering neighbourhood partitions from geo-tagged natural language descriptions of 40,346 London Airbnb listings. We embed descriptions using an Large Language Model-based embedding model and construct a weighted kNN graph that integrates geographic proximity and semantic similarity between properties. Leiden community detection on this graph yields spatially contiguous neighbourhood partitions, which we validate against three indicators of urban structure: functional and commercial concentration via amenity distribution, accessibility and urban connectivity via transit structure, and social composition via socio-economic patterning. While these indicators are not an exhaustive representation of urban characteristics, they do provide an interpretable basis against which our spatio-semantic partitions can be assessed. Communities align strongly with amenity structure, with POI density higher in community cores than peripheries in 91.9% of cases, and align moderately with socio-economic structure (global NMI = 0.193). We also demonstrate qualitative alignment between transit structure and identified partitions. An ablation study shows that semantic information improves amenity alignment substantially more than socio-economic alignment, consistent with the leisure- and tourism-oriented content of listing descriptions.

Submission Type. Analysis, Model.

BoK Concepts. [GC3-8-1] Natural language processing, [CF2-1] Perception and Cognition of Geographic Phenomena, [CF2-4] Place and landscape

Keywords. Neighbourhoods, Natural Language Processing, Graphs, Large Language Models

1 Introduction

Urban neighbourhoods reflect both objective and subjective phenomena. Neighbourhoods range from administrative divisions to more subjective discretisations of urban space shaped by amenity access, transit connectivity, and socio-economic composition. (Martí et al., 2022; Galster, 2001). Shared use of place names is a clear initiator of neighbourhood structure (Brindley et al., 2014; McKenzie et al., 2018), however, more subtle linguistic indicators can signal perceived proximity (Shingleton and Basiri, 2025) and shared local context (Stock et al., 2022) - suggesting a need for a semantic understanding of geographic text which goes beyond grouping of toponyms. In this paper, we examine how geography-rich natural language can be used to identify spatio-semantic neighbourhoods in London, UK, and what these partitions reveal about the drivers of neighbourhood structure within the context of the short-term rental market.

The proliferation of geographically rich text in online sources, along with the development of advanced Natural Language Processing techniques, has enabled data-driven approaches to the delineation of urban neighbourhoods. Brindley et al. (2014) showed that extraction and mapping of place names from online textual data sources can reflect expected urban neighbourhood distributions. McKenzie et al. (2018) applied a random forest model to N-grams within user-contributed property listings in Washington, DC, Seattle, WA, and Montreal, QC; identifying place-name clusters reflective of urban

neighbourhoods. Poorthuis (2018) used natural language processing over social media text to derive dynamic neighbourhoods, demonstrating that place-based language can encode spatial structure. More recently, Shingleton and Basiri (2025) showed that Airbnb descriptions contain semantic clues about perceived geographic proximity.

By using shared toponym usage as an indicator of neighbourhood membership, the approaches described above fail to account for wider semantic context in which people describe their lived environments. In this paper, we combine large language model (LLM)-derived semantic embeddings with an explicit spatial graph to recover contiguous, interpretable neighbourhood partitions and test what urban factors they align with.

We embed Airbnb listing descriptions using the E5-Mistral-7B-Instruct text embedding model (Wang et al., 2024) to obtain dense semantic representations of listing text. We then construct a weighted spatio-semantic graph in which each listing is connected to both its nearest geographic and semantic neighbours, with edge weights combining semantic similarity and geographic proximity. We apply Leiden community detection (Traag et al., 2019) to this graph to obtain neighbourhood partitions. We assess the degree to which our identified neighbourhoods conform to other indicators of urban structure, including POI density, transit networks, and socio-economic patterning.

We show that the resulting neighbourhoods are spatially structured around amenity and transit access, and are partially reflective of socio-economic structure within London. This work makes three contributions:

- **Methodological novelty**: We combine LLM embeddings with an explicitly spatial graph and perform community detection to identify semantic neighbourhoods. We use ablation to isolate and evaluate the influence of semantic information, compared to a purely spatial baseline.
- **Empirical novelty**: The methods and data used present a scalable means of identifying neighbourhood structure in large urban areas, and show that this structure aligns with other urban neighbourhood indicators.
- **Conceptual novelty**: The paper moves beyond named-place extraction and toward semantically coherent, contiguous neighbourhoods defined by descriptive language and local characteristics.

2 Methods

Our analysis combines geographic and semantic information from Airbnb listing descriptions to derive neighbourhood-level communities across London. Semantic embeddings are used to construct a spatio-semantic graph of properties, over which coherent communities are identified via Leiden community detection (Traag et al., 2019).

2.1 Data Collection and Preparation

We use Airbnb property descriptions from London UK, extracted from the Inside Airbnb dataset (Murray Cox, 2025). The descriptions often provide detailed information about both the property itself, as well as the surrounding neighbourhood. The dataset contains a total of 93 182 property descriptions, scraped between 06/09 and 11/09 2024. After deduplication and removal of properties with very short descriptions (fewer than 250 characters), a total of 40 346 descriptions were retained for further analysis.

2.2 Semantic Embeddings

We embed each Airbnb listing description using the E5-Mistral-7B embedding model (Wang et al., 2024), producing a vector $x_i \in \mathbb{R}^{4096}$ for listing i. Embeddings are extracted using last-token pooling over the final hidden layer and are ℓ_2-normalised prior to downstream analysis (Wang et al., 2024).

The resulting embeddings provide a dense representation of the semantic content of each listing description, so that listings described with similar language tend to be closer in embedding space. Prior work has shown that such embeddings can encode geographic signals (Shingleton et al., 2026), motivating their use in spatio-semantic graph construction.

2.3 Graph Construction

We construct an undirected weighted spatio-semantic graph $G = (V, E, W)$ over our listings, with one vertex per property. We add an edge (i, j) if listing j lies among the k_g nearest neighbours of i in geographic space (haversine distance), or k_s nearest neighbours of i in semantic space (cosine similarity of embeddings).

Each edge $\{i, j\}$ is assigned weight

$$w_{i,j} = \tilde{s}_{i,j} \cdot \exp\left(-\frac{d_{i,j}}{T}\right), \tag{1}$$

where $d_{i,j}$ is the haversine distance between properties i and j and T controls distance decay. Let $s_{i,j}$ denote cosine similarity between embedding vectors x_i and x_j. Cosine similarities are clipped so that $s'_{i,j} = \max(0, s_{i,j})$, before being min-max scaled within each node's neighbourhood $\tilde{s}_{i,j} \in [\varepsilon, 1]$, where $\varepsilon = 10^{-4}$ acts as a floor to prevent zero-weighted edges. Weights are symmetrised by $w_{i,j} \leftarrow \max(w_{i,j}, w_{j,i})$.

Neighbourhood min–max scaling avoids forming isolated clusters in areas where listings have uniformly low semantic similarity to the rest of the dataset, and increases weight diversity within semantically homogenous neighbourhoods. Because the lowest similarities in a local

neighbour set typically arise from geographically close but semantically weak edges, local scaling uses these to anchor the lower bound, helping prevent the weighting from effectively collapsing to a distance-only term.

2.4 Leiden Community Detection

We identify communities in the spatio-semantic graph using the Leiden algorithm (Traag et al., 2019), which optimises modularity and produces well connected communities (Liu et al., 2025). We apply Leiden to the undirected weighted graph using the `igraph` implementation (Csárdi and Nepusz, 2006). We report summary statistics for Leiden communities generated across a range of values for k_g, k_s and T.

2.5 Points of Interest and Community Structure

Access to local amenities often shapes neighbourhood structure. To assess whether our spatio-semantic neighbourhoods align with this expectation, we compare Leiden communities against the distribution of points-of-interest (POIs) in London. We use food and drink-related POIs as a proxy for neighbourhood amenity density: these venues are abundant, geographically widespread, strongly co-located with other commercial and leisure activity, and are frequently referenced in Airbnb property descriptions. While food and drink venues do not capture all amenity types, this provides a stable and interpretable proxy for amenity concentration.

POI data are extracted from the Geographic Data Service (Berragan, 2026). Of 347,383 POIs within the Greater London boundary, we retain 29,216 whose main category field contains one of the substrings *restaurant*, *cafe*, *pub*, or *bar*.

For each Leiden community, we assess whether POI accessibility differs between its geographic core and periphery. We define the community centre as the medoid of constituent property locations Kaufman and Rousseeuw (1990). We then define *core* properties as the 10% closest to the medoid and *periphery* properties as the 10% furthest from the medoid.

For each property, we compute the distance to its $k = 20$ nearest POIs. Let d_{core} and d_{peri} denote the median of these distances over core and periphery properties respectively, and let d_{all} denote the corresponding median over all properties in the community. We report the normalised core–periphery contrast:

$$\Delta_d = \frac{d_{\text{peri}} - d_{\text{core}}}{d_{\text{all}}}. \tag{2}$$

Positive values indicate that core properties have smaller POI distances than peripheral properties (i.e., higher amenity density around community centres), consistent with communities reflecting amenity structure.

2.6 Socio-Economic Groups and Community Structure

To assess the degree to which our spatio-semantic partitions align with socio-economic factors, we compare our derived partitions against the Output Area Classification (OAC) of Singleton and Longley (2024). Using 2021 UK Census variables, they derive socio-economic classifications for Output Areas (OAs) in London (Office for National Statistics). Each OA is assigned to one of seven *supergroups*, which are further differentiated into *groups* and *subgroups*. For example, OA *E00019467* in Richmond upon Thames has classification *5a3*, corresponding to supergroup 5, group 5a, and subgroup 5a3.

For each Airbnb listing, we identify the socio-economic label at the property location, yielding a categorical label $y_i \in \{1, 2, \ldots, K\}$ for listing i. To reduce fine-scale spatial noise in the socio-economic labelling, we evaluate a denoised variant in which labels are aggregated to an H3 grid (resolution 9) (Bousquin, 2021). For each H3 cell, we assign the dominant socio-economic class by area-weighted intersection between the hex cell and the census polygons; listings in that cell then inherit the hex-level label.

For each listing i, we compare three alternative spatial partitions:

- **Leiden communities** — derived from the spatio-semantic graph,
- **K-means partition** — derived from semantic embeddings only,
- **Administrative partition** — London's 33 boroughs.

2.6.1 Entropy of socio-economic labels within partitions

For a given partition into parts $c \in \mathcal{C}$, we quantify the diversity of socio-economic labels within each part using Shannon entropy (Shannon, 1948).

For a partition c, let $p_{c,l}$ be the proportion of listings in c assigned to socio-economic class l. We compute the *normalised* entropy:

$$H(c) = -\frac{\sum_{l=1}^{L} p_{c,l} \log p_{c,l}}{\log L}. \tag{3}$$

We summarise performance across parts using the size-weighted mean:

$$\bar{H} = \frac{\sum_{c \in \mathcal{C}} n_c \, H(c)}{\sum_{c \in \mathcal{C}} n_c}, \tag{4}$$

where n_c is the number of listings in part c. Lower values indicate parts that are more homogeneous with respect to socio-economic labels.

2.6.2 Global Normalised Mutual Information

To quantify overall agreement between a partition C and socio-economic labels Y, we compute normalised mutual information (NMI):

$$\mathrm{NMI}(Y,C) = \frac{I(Y;C)}{\frac{1}{2}\left(H(Y)+H(C)\right)}, \tag{5}$$

where $I(\cdot;\cdot)$ is mutual information and $H(\cdot)$ is Shannon entropy (Kvålseth, 2017). $\mathrm{NMI} \in [0,1]$, where 0 indicates no association and 1 indicates perfect agreement.

2.6.3 Within-borough NMI

Because boroughs are large and internally heterogeneous, we also test whether Leiden and K-means provide meaningful sub-borough structure. We compute within-borough NMI by comparing socio-economic labels y_i to the composite label (b_i, c_i), where b_i is the borough of listing i and c_i is its partition label:

$$\mathrm{NMI}_w = \mathrm{NMI}(Y, (B \times C)). \tag{6}$$

We compare NMI_w to a within-borough permutation null (partition labels shuffled within boroughs), reporting the null mean and standard deviation over 1000 permutations.

2.7 Ablation and Robustness

To test whether adding semantic information improves performance compared to a spatial-only graph, we run an ablation over graph construction and weighting.

We test three variants

1. spatial only version with $k_g = 23$ geo-kNN edges, with weights $w_{i,j} = \exp(-d_{i,j}/T)$ and $T = 1.0$;
2. a spatial graph with $k_g = 23$ geo-kNN edges, using semantic-only weights $w_{i,j} = \tilde{s}^{\tau}_{i,j}$ and $\tau = 1.0$;
3. a full spatio-semantic graph with $k_g = 10$ geo-kNN neighbours, $k_s = 6$ semantic-kNN neighbours, and weights as described in eq 1, with $T = 1.0$.

Parameters were selected to control for community count (≈ 60 communities) to avoid trivial improvements from over-fragmentation. Leiden community detection was repeated 1000 times per setting. We report the mean and standard deviation across five metrics: size-weighted mean normalised entropy, global NMI, within-borough NMI, and the proportions of communities with $\Delta_d > 0$ and $\Delta_d > 0.2$.

2.8 Data Availability and Reproducibility

Code and derived data are available in an Open Science Foundation repository[1], including environment and hardware specifications and scripts to reproduce all results. As embedding inference may vary across hardware, we also provide the embedding vectors used in this analysis.

[1] https://osf.io/bnh8v

3 Results

3.1 Parameter Selection

Figure 1 reports a grid search over $k = k_g = k_s$ and the distance-temperature T, minimum community size, and socio-economic validation.

Low T and low k produce many small communities with high modularity. These settings also tend to inflate socio-economic validation metrics, but yield less interpretable and less robust partitions. A stable regime emerges for $5 \leq k \leq 10$ and $0.8 \leq T \leq 1.5$, where community counts remain ∼50–70 with similar validation performance.

Figure 2 varies k_g and k_s at fixed $T = 1.0$. Small k_g or k_s again produces fragmented communities, with a broad plateau for $5 \leq k_g, k_s \leq 10$ yielding 55-62 communities and $\mathrm{NMI}_g \approx 0.193$. We select $k_g = 10$, $k_s = 6$ and $T = 1.0$ as a representative point on this plateau, producing 61 communities with a minimum size of 102, and modularity 0.949.

3.2 Spatial distribution of communities

Figure 3 maps the Leiden communities identified in the spatio-semantic graph and a semantic-only baseline obtained by K-means clustering of description embeddings ($K = 61$, matching the Leiden community count). The K-means map shows that the text embeddings already encode substantial spatial structure. Leiden communities are more spatially contiguous, consistent with the additional geographic information introduced by the graph construction.

3.3 Community structure and POI distribution

Figure 4 illustrates four boroughs (Southwark, Lambeth, Hackney, and Camden) that contain multiple Leiden communities. Community boundaries often coincide with locally lower POI density, suggesting that amenity access contributes to community structure. Figure 5 summarises this pattern across all communities using $\Delta_d = (d_{peri} - d_{core})/d_{all}$. We find $\Delta_d > 0$ in 91.8% of communities, and $\Delta_d > 0.2$ in 82.0% of communities.

3.4 Community structure and transit access

In many cases, communities appear to structure around transit routes. Figure 6 overlays London Underground and Overground routes (Transport for London, 2026) on Leiden communities (aggregated to H3, resolution 9).

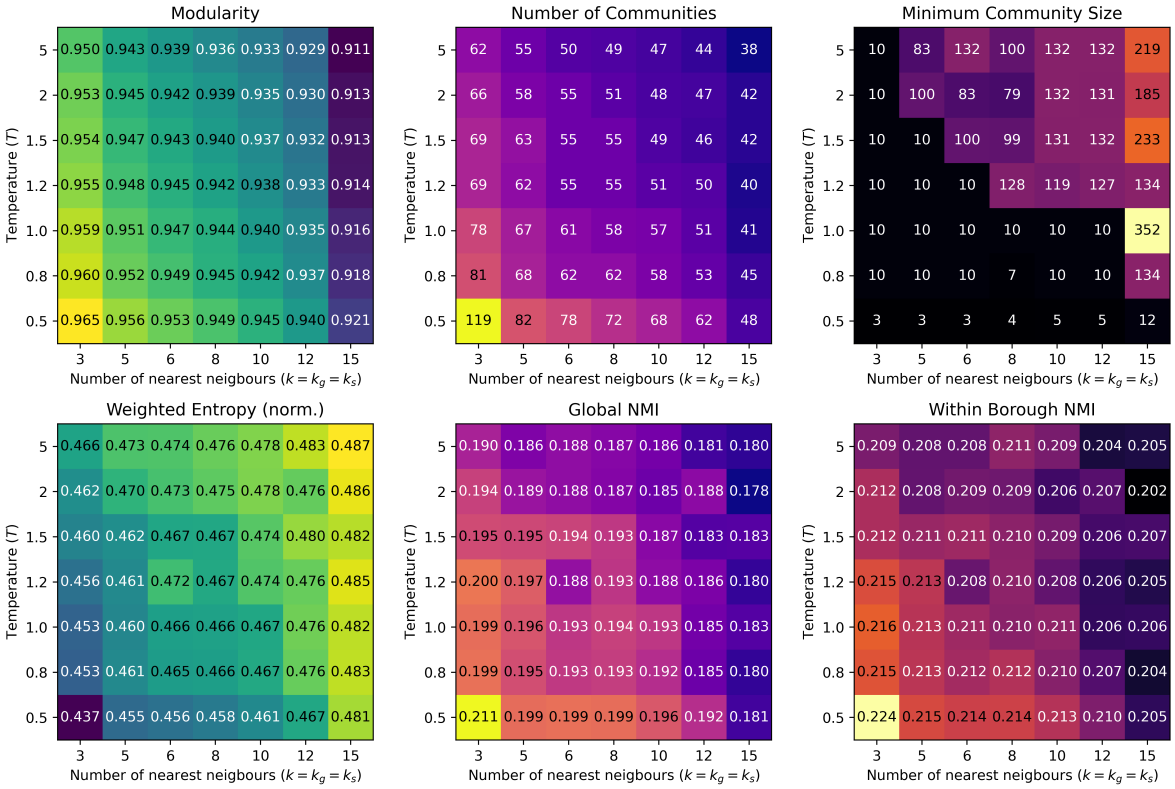

Figure 1. Grid search results across $k = k_g = k_s$ and T.

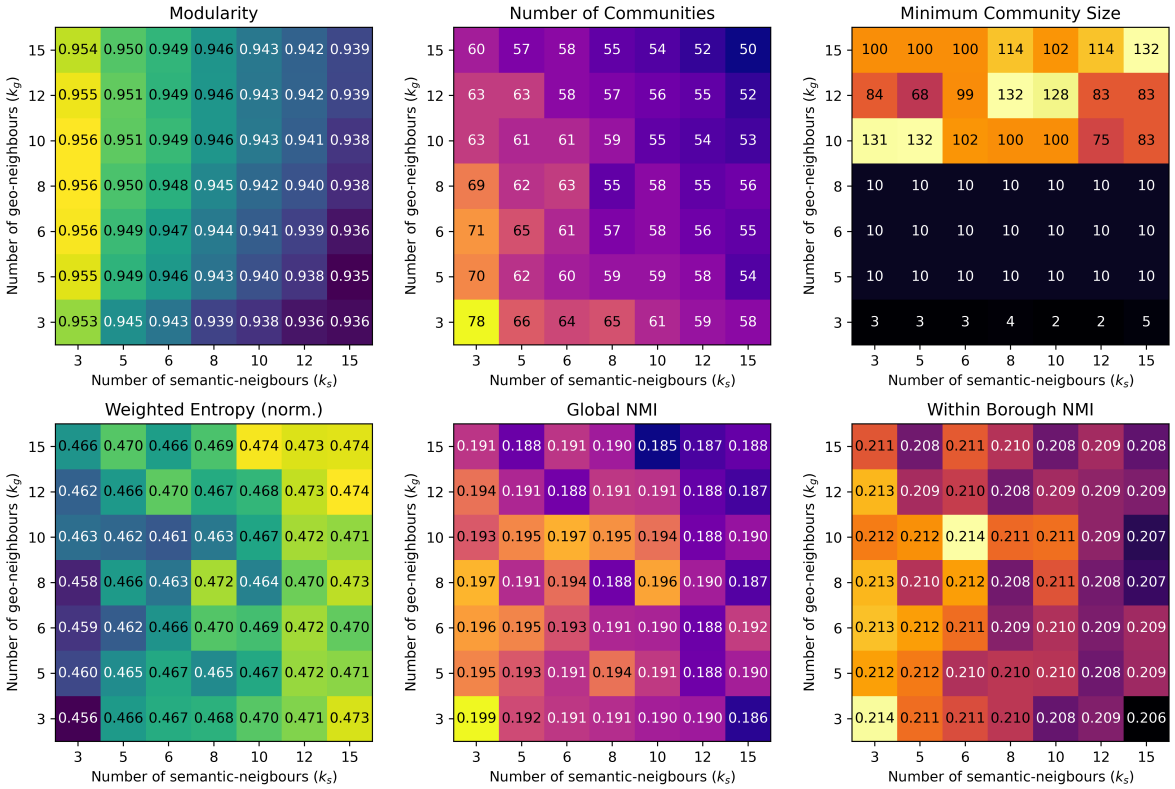

Figure 2. Grid search results across graph constructions using k_g geographic neighbours and k_s semantic neighbours at fixed distance-weighting temperature $T = 1.0$.

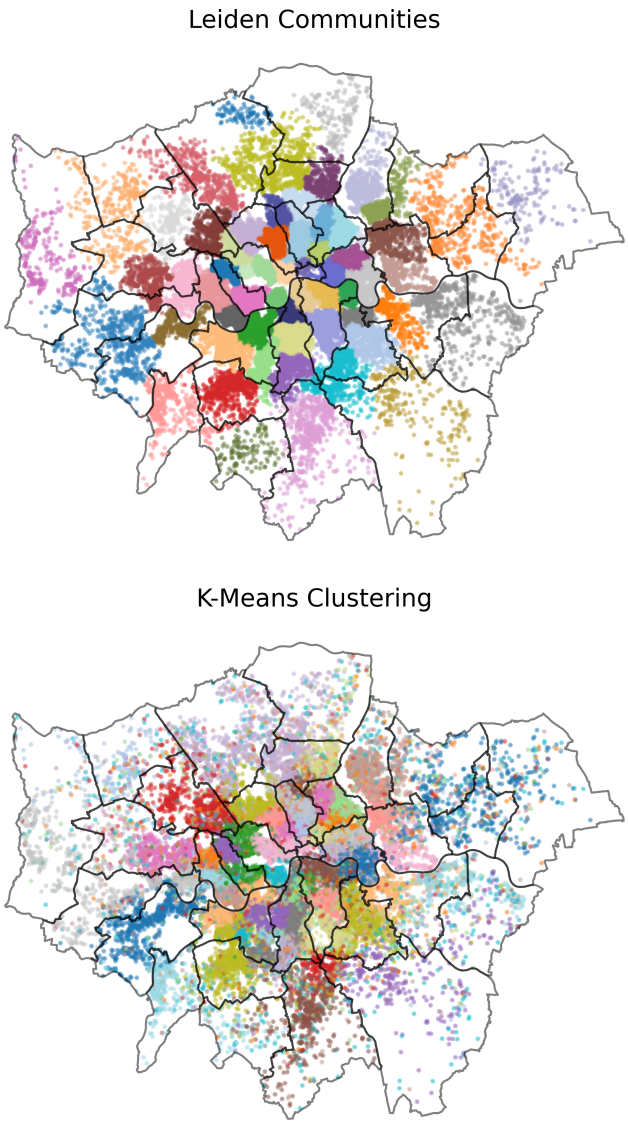

Figure 3. Distribution of Airbnb property communities identified via Leiden community detection on a spatio-semantic graph (top) and K-means clustering on the property description embeddings (bottom, $K = 61$).

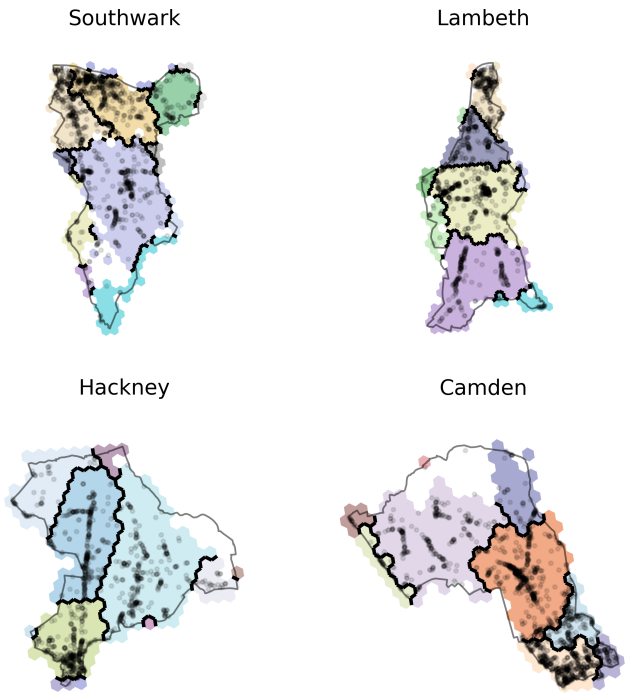

Figure 4. Leiden communities in Southwark, Lambeth, Hackney, and Camden shown as coloured regions, with food and drink POIs shown as black points. Communities are aggregated to H3 (resolution 9) for visualisation, assigning each hex the modal community label among listings within the cell.

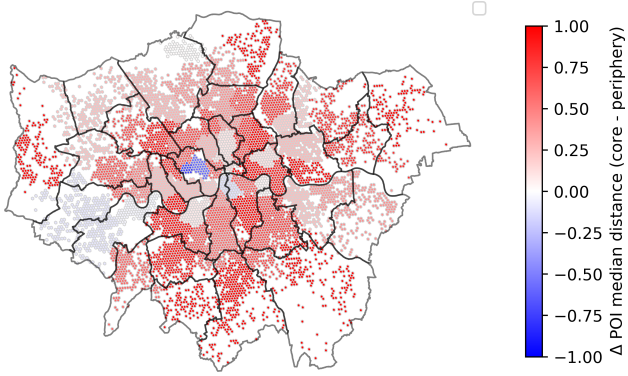

Figure 5. Community-level amenity gradient, visualised on an H3 grid (resolution 9). Colour indicates $\Delta_d = (d_{peri} - d_{core})/d_{all}$, based on median distance to $k = 20$ nearest POIs. Red indicates lower POI density toward community edges.

Qualitatively, transit lines often pass through community interiors rather than along their boundaries.

3.5 Community structure and socio-economic group

We assess the extent to which Leiden communities align with socio-economic structure using the OAC socio-economic subgroups. Table 1 reports the weighted normalised Shannon entropy ($\bar{H}$), global NMI (NMI_g), and within-borough NMI (NMI_w) for three partitions: boroughs, Leiden communities, and K-means partitions of the embedding space. we also report a null baseline obtained by randomly permuting community labels within each borough over 1000 iterations. Overall, Leiden communities achieve the lowest entropy and highest global NMI, suggesting that the spatio-semantic partition captures socio-economic structure more effectively than borough boundaries or semantic-only K-means partitions.

Figure 7 compares Leiden communities against dominant OAC subgroups (hex-aggregated at H3 resolution 9) for four boroughs (Brent, Ealing, Islington, Lambeth). In Brent and Ealing, Leiden communities align well with socio-economic variation, reflected by relatively high NMI_w (0.271 and 0.211). Islington is comparatively homogeneous (dominated by subgroup 3c2), and correspondingly NMI_w is lower (0.148), implying that other neighbourhood drivers (e.g., amenities or transit) may dominate partitioning. Lambeth highlights a

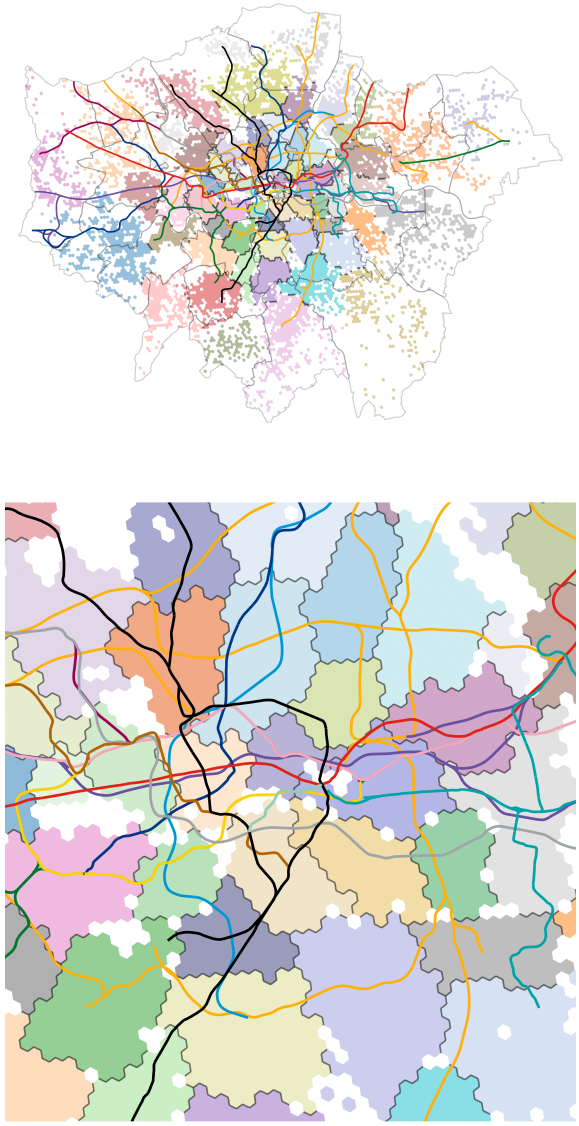

Figure 6. Community structure across London overlaid with the London Underground and London Overground transit routes. Colours indicate different routes, and correspond to the standard colouring used on the London Underground Map (Transport for London, 2025). The lower map shows the detail within the central part of London.

limitation of NMI-based validation: Leiden isolates both homogeneous and heterogeneous regions with apparent spatial coherence, but the mixed regions are penalised by the metric ($\mathrm{NMI}_w = 0.121$).

3.6 Ablation: contribution of semantic information

To test whether semantic information improves community structure beyond a spatial-only graph, we compared three graph constructions and weighting schemes (Section 2.7), with parameters chosen to control for community count (≈ 60 communities) and avoid trivial gains from over-fragmentation. For each setting, Leiden community detection was repeated 1000 times on a fixed graph, and we report mean $\pm$ standard deviation across runs. As expected, spatial effects alone account for a large portion of our community structure, however the contribution from semantic information remains significant within some contexts.

Table 1. Socio-economic alignment metrics for alternative partitions. All results use the hex-denoised subgroup unless stated otherwise.

Partition	Metric		
	$\bar{H}$	NMI_g	NMI_w
Borough	0.498	0.177	-
Leiden communities	0.461	0.197	0.214
K-means clusters	0.506	0.143	0.189
Random partitioning	-	-	0.148 $\pm$0.002

Across socio-economic validation metrics, differences between constructions were modest. All three settings achieved similar normalised entropy ($\bar{H}$) and global NMI (NMI_g), with the spatio-semantic graph yielding a small but consistent improvement in NMI_g relative to spatial-only baselines (Table 2). Within-borough NMI (NMI_w) was effectively saturated across settings, suggesting that most of the socio-economic signal is already captured by geographic adjacency at this scale and label resolution.

In contrast, semantic information produced a clearer improvement in amenity alignment compared to the spatial-only graph. Using the POI distance gradient metric Δ_d (Section 2.5), the spatio-semantic construction increased the proportion of communities with higher POI density at the core ($\Delta_d > 0$), and also increased the proportion with a substantial core–periphery difference ($\Delta_d > 0.2$), relative to both spatial-only variants. Hence, although the semantic information in Airbnb descriptions adds limited incremental value for explaining socio-economic partitions, it contributes more strongly to capturing neighbourhood structure related to amenity concentration. This is consistent with Airbnb descriptions encoding signals that are relevant to the short-term rental market, such as proximity to leisure amenities.

4 Discussion

This study presents a scalable and adaptable workflow for recovering perceived neighbourhood structure from unstructured geo-tagged text, identifying spatio-semantic neighbourhoods that align with several indicators of neighbourhood structure. By augmenting a spatial graph with semantic embeddings of geographically-rich Airbnb descriptions, we capture the contiguous and interpretable nature of neighbourhood partitions. A key finding from the ablation study is that, while spatial

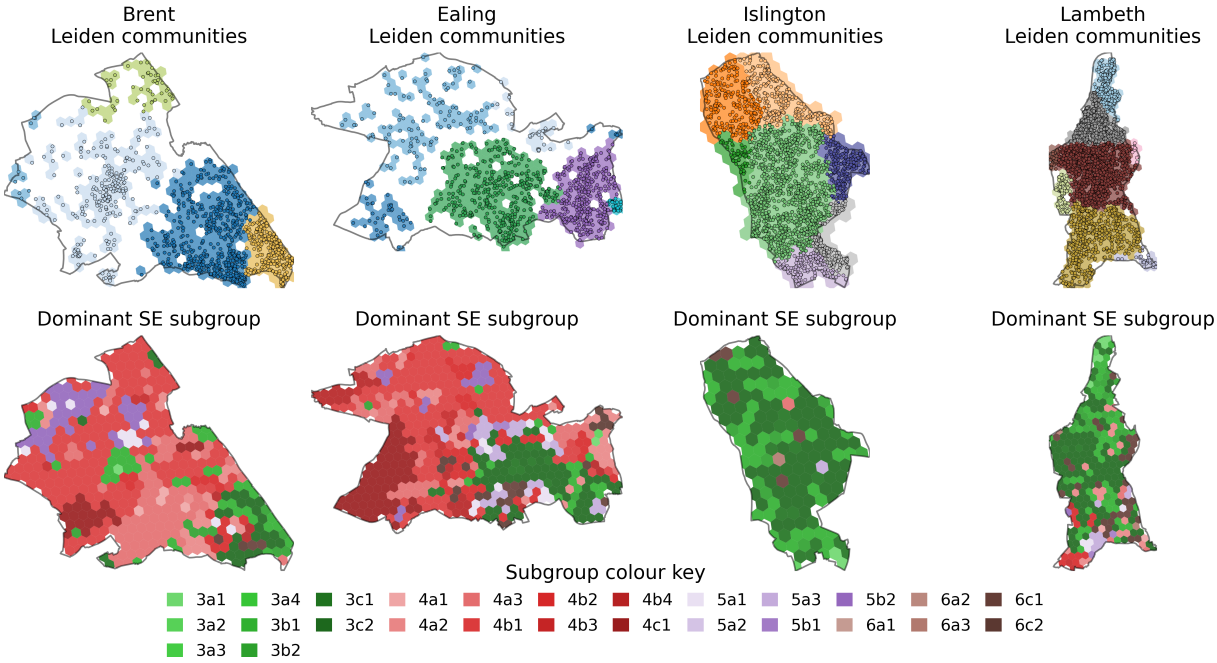

Figure 7. A comparison of the identified spatio-semantic Leiden communities and the socio-economic subgroups aggregated to an H3 grid (resolution=9).

Table 2. Ablation study over graph/weight variants (mean ± SD over 1000 Leiden runs). S–D: spatial graph + distance weights; S–S: spatial graph + semantic weights; SS–M: spatio-semantic graph + mixed weights. KS p compares S–D vs SS–M (two-sample Kolmogorov–Smirnov). Parameters were chosen to control for number of Leiden communities (≈ 60) to avoid trivial improvement from over fragmentation.

Metric	S–D	S–S	SS–M	KS p (S–D vs SS–M)
# Communities	60.4 ± 1.566	60.4 ± 1.453	60.6 ± 1.417	0.097
Norm. entropy ($\bar{H}$)	0.465 ± 0.002	0.464 ± 0.002	0.464 ± 0.002	< 0.001
Global NMI (NMI_g)	0.191 ± 0.002	0.192 ± 0.002	0.194 ± 0.002	< 0.001
Within borough NMI (NMI_w)	0.213 ± 0.002	0.213 ± 0.001	0.212 ± 0.001	< 0.001
Positive POI gradient ($P(\Delta_d > 0)$)	0.781 ± 0.036	0.812 ± 0.033	0.901 ± 0.020	< 0.001
Significant POI gradient ($P(\Delta_d > 0.2)$)	0.685 ± 0.040	0.694 ± 0.038	0.812 ± 0.032	< 0.001

adjacency captures baseline socio-economic patterns, the inclusion of semantic information is a primary driver for distinguishing functional sub-areas, particularly regarding amenity access. This suggests that Airbnb descriptions effectively encode a functional and amenity-focused geography.

However, it is important to acknowledge that the spatio-semantic neighbourhoods derived here are based on Airbnb descriptions, which naturally reflect a "tourist gaze", a representation of the city curated to attract short-term visitors. Consequently, while our method produces coherent partitions, these should be interpreted as representations of the "short-term rental city" rather than a holistic view of urban residence. Future work should therefore extend this approach to alternative datasets representing different urban perspectives, allowing for comparative analyses that deepen our understanding of how different populations perceive and segment urban areas.

In conclusion, this study establishes a novel framework for identifying perceived neighbourhood structure through semantic embeddings derived from LLMs. We demonstrate that by leveraging unstructured Airbnb property descriptions, it is possible to define spatially contiguous and functionally coherent neighbourhood boundaries. These findings validate the potential of LLMs as powerful tools for urban analysis, beyond simply generating geographically relevant text, and offering a scalable method to map cities through specific perceptual lenses. As the volume of unstructured text and multimodal urban data continues to grow, such semantic-first approaches will become increasingly critical for decoding the complex and subjective layers of the urban fabric.

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. Specifically, the AI tools were utilized for improving grammar and sentence structure but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

Acknowledgements

The authors acknowledge the support from the UK Research and Innovation Future Leaders Fellowships "Indicative Data: Extracting 3D Models of Cities from Unavailability and Degradation of Global Navigation Satellite Systems (GNSS)" [MR/S01795X/2] and "Missing Data as Useful Data" [MR/Y011856/1], AI Metascience Fellowship "Generative AI and the Future of Research Software Engineering" [UKRI2599], and the Alan Turing Institute-DSO partnership project on "Multi-Lingual and Multi-Modal Location Information Extraction".

Author Contributions

JS conceptualised the project, completed the data curation, initial experimental design, formal analysis, creation of research software, and initial drafting of the manuscript. YSB contributed to ideation and development of the experimental and analytic methods, and initial drafting of the manuscript. AB, YW, and MW contributed to ideation and development of the experimental and analytic methods, and edited the original draft of the manuscript.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Berragan, C.: Point of Interest data for the United Kingdom, Geographic Data Service (GeoDS) dataset, version 1.1, https://data.geods.ac.uk/dataset/point-of-interest-data-for-the-united-kingdom, last updated 2026-01-15. Overture release 2024-09-18.0. Licensed under CDLA-Permissive-2.0. Accessed 2026-02-16., 2026.

Bousquin, J.: Discrete Global Grid Systems as scalable geospatial frameworks for characterizing coastal environments, Environmental Modelling Software, 146, 105 210, https://doi.org/https://doi.org/10.1016/j.envsoft.2021.105210, 2021.

Brindley, P., Goulding, J., and Wilson, M. L.: A data driven approach to mapping urban neighbourhoods, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14, p. 437–440, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2666310.2666473, 2014.

Csárdi, G. and Nepusz, T.: The igraph software package for complex network research, in: Complex Systems 2006, InterJournal, https://api.semanticscholar.org/CorpusID:16923281, 2006.

Galster, G.: On the Nature of Neighbourhood, Urban Studies, 38, 2111–2124, https://doi.org/10.1080/00420980120087072, 2001.

Kaufman, L. and Rousseeuw, P. J.: Partitioning Around Medoids (Program PAM), chap. 2, pp. 68–125, John Wiley Sons, Ltd, https://doi.org/https://doi.org/10.1002/9780470316801.ch2, 1990.

Kvålseth, T. O.: On Normalized Mutual Information: Measure Derivations and Properties, Entropy, 19, https://doi.org/10.3390/e19110631, 2017.

Liu, W., Cai, H., Xing, H., Hu, S., Tan, Z., and Song, C.: Alleviating the resolution limit problem in spatial community detection: a local network structure-based method, International Journal of Geographical Information Science, 39, 510–532, https://doi.org/10.1080/13658816.2024.2421778, 2025.

Martí, P., Serrano-Estrada, L., Nolasco-Cirugeda, A., and Baeza, J. L.: Revisiting the Spatial Definition of Neighborhood Boundaries: Functional Clusters versus Administrative Neighborhoods, Journal of Urban Technology, 29, 73–94, https://doi.org/10.1080/10630732.2021.1930837, 2022.

McKenzie, G., Liu, Z., Hu, Y., and Lee, M.: Identifying Urban Neighborhood Names through User-Contributed Online Property Listings, ISPRS International Journal of Geo-Information, 7, https://doi.org/10.3390/ijgi7100388, 2018.

Murray Cox, John Morris, T. H.: Inside Airbnb, https://insideairbnb.com/about/, 2025.

Office for National Statistics: Statistical geographies, https://www.ons.gov.uk/methodology/geography/ukgeographies/statisticalgeographies, oNS Methodology: UK geographies.

Poorthuis, A.: How to Draw a Neighborhood? The Potential of Big Data, Regionalization, and Community Detection for Understanding the Heterogeneous Nature of Urban Neighborhoods, Geographical Analysis, 50, 182–203, https://doi.org/https://doi.org/10.1111/gean.12143, 2018.

Shannon, C. E.: A mathematical theory of communication, The Bell System Technical Journal, 27, 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x, 1948.

Shingleton, J. and Basiri, A.: How close is "close"? An analysis of the spatial characteristics of perceived proximity using Large Language Models, AGILE: GIScience Series, 6, 11, https://doi.org/10.5194/agile-giss-6-11-2025, 2025.

Shingleton, J., Bıçakçı, Y. S., Wang, Y., and Basiri, A.: Emergent Spatio-Semantic Structure in Large Language Model Embedding Spaces, https://doi.org/10.1080/17489725.2025.2594193, 2026.

Singleton, A. D. and Longley, P. A.: Classifying and mapping residential structure through the London Output Area Classification, Environment and Planning B: Urban Analytics and City Science, 51, 1153–1164, https://doi.org/10.1177/23998083241242913, 2024.

Stock, K., Jones, C. B., Russell, S., Radke, M., Das, P., and Aflaki, N.: Detecting geospatial location descriptions in natural language text, International Journal of Geographical Information Science, 36, 547–584, https://doi.org/10.1080/13658816.2021.1987441, 2022.

Traag, V. A., Waltman, L., and Van Eck, N. J.: From Louvain to Leiden: guaranteeing well-connected communities, Scientific Reports, 9, 5233, https://doi.org/10.1038/s41598-019-41695-z, 2019.

Transport for London: TfL Colour Standard, https://content.tfl.gov.uk/tfl-colour-standard.pdf, 2025.

Transport for London: TfL GIS Open Data Hub, https://gis-tfl.opendata.arcgis.com/, 2026.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F.: Improving Text Embeddings with Large Language Models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), edited by Ku, L.-W., Martins, A., and Srikumar, V., pp. 11 897–11 916, Association for Computational Linguistics, Bangkok, Thailand, https://doi.org/10.18653/v1/2024.acl-long.642, 2024.