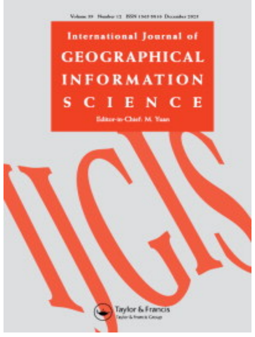


International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

A graph neural network for small-area estimation: integrating spatial regularisation, heterogeneous spatial units, and Bayesian inference

Pengyuan Liu, Yang Chen, Xiucheng Liang, Hao Li, Filip Biljecki & Rudi Stouffs

To cite this article: Pengyuan Liu, Yang Chen, Xiucheng Liang, Hao Li, Filip Biljecki & Rudi Stouffs (29 Dec 2025): A graph neural network for small-area estimation: integrating spatial regularisation, heterogeneous spatial units, and Bayesian inference, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2597971

To link to this article: https://doi.org/10.1080/13658816.2025.2597971

Published online: 29 Dec 2025.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2597971

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

A graph neural network for small-area estimation: integrating spatial regularisation, heterogeneous spatial units, and Bayesian inference

Pengyuan Liu[a,b], Yang Chen[c,d], Xiucheng Liang[c], Hao Li[e], Filip Biljecki[c,f] and Rudi Stouffs[c]

[a]Urban Analytics Subject Group, Urban Studies & Social Policy Division, University of Glasgow, Glasgow, UK; [b]Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, Glasgow, UK; [c]Department of Architecture, National University of Singapore, Singapore, Singapore; [d]School of Geography, Nanjing Normal University, Nanjing, China; [e]Department of Geography, National University of Singapore, Singapore, Singapore; [f]Department of Real Estate, National University of Singapore, Singapore, Singapore

ABSTRACT
Fine-resolution spatial analytics are essential for urban planning and policy-making, yet traditional small-area estimation often struggles with sparse, hierarchical, or imbalanced data. This paper introduces a Spatially Regularised Bayesian Heterogeneous Graph Neural Network (SR-BHGNN) that integrates multiple census tract levels within a unified framework. The model builds a heterogeneous graph where nodes represent spatial units at different scales, edges encode adjacency or membership, and Bayesian inference quantifies uncertainty in parameters and predictions. A spatial regularisation term, inspired by Tobler's First Law of Geography, penalises large discrepancies between neighbouring nodes, reducing errors in imbalanced datasets and ensuring coherent local estimates. We evaluate SR-BHGNN through two London case studies, population estimation and PM $_{2.5}$ prediction, comparing it against random forests, single-level GNNs, and spatial hierarchical Bayesian estimation. SR-BHGNN achieves strong performance gains, with classification accuracies of 0.85 for population estimation and 0.81 for PM $_{2.5}$ prediction. Its Bayesian design produces posterior distributions that capture uncertainty, enabling policy-relevant insights into vulnerable neighbourhoods or priority intervention zones (e.g. low-emission areas). These results demonstrate that SR-BHGNN advances the state of the art in small-area estimation, offering a flexible, uncertainty-aware framework for diverse urban analytics applications.

ARTICLE HISTORY
Received 7 August 2025
Accepted 27 November 2025

KEYWORDS
Small-area estimation; spatial regularisation; uncertainty quantification; Tobler's first law

1. Introduction

Urban spaces are inherently hierarchical, with spatial data often spanning multiple geography scales and varying in granularity and quality (Mu and Wang 2006,

CONTACT Pengyuan Liu Pengyuan.Liu@glasgow.ac.uk

Fotheringham et al. 2009, Anselin 2013). Fine-scale spatial information is essential for informed local decision-making, planning, and policy implementation, yet these data are frequently unavailable, incomplete, or costly to collect. (Tate and Atkinson 2001, Goodchild 2007). Conversely, data aggregated at coarser administrative or geographic scales tend to be more readily accessible and reliable but lack sufficient spatial detail to support precise local interventions (Wang *et al.* 2024). Such a scale mismatch is a persistent challenge in urban analytics, motivating researchers to explore approaches that can accurately estimate fine-scale information by 'borrowing strength' from coarser-resolution datasets and spatial patterns (Ghosh and Rao 1994, Pfeffermann 2002, Rao and Molina 2015, Luo *et al.* 2025).

Traditionally, statistical methods such as small-area estimation (SAE) (Rao and Molina 2015, Whitworth *et al.* 2017) and geographically weighted regression (GWR) (Fotheringham *et al.* 2009) have been widely adopted for such downscaling tasks. SAE methods typically rely on hierarchical Bayesian or empirical Bayes frameworks that introduce random effects or partial pooling to transfer information across administrative units, particularly in cases where local data are sparse (Guan *et al.* 2011, Rao and Molina 2015, Yao *et al.* 2017). Hierarchical models explicitly represent both local variations and broader-scale commonalities, acknowledging that smaller areas are nested within larger geographic entities (Banerjee *et al.* 2003, Cressie 2015). However, these statistical methods generally assume linear relationships or strong parametric spatial structures, restricting their effectiveness when applied to complex urban environments characterised by heterogeneous data and non-linear spatial dependencies.

Recent advancements in machine learning and spatial data science offer more flexible solutions for addressing these complexities (Singleton *et al.* 2020, Jiang and Rao 2020). Graph Neural Networks (GNNs), in particular, have gained prominence for modelling spatial interactions, as they naturally represent spatial units as nodes and spatial relationships as edges within a network structure (Liu and Biljecki 2022, Wang and Zhu 2024, Zhu and Ma 2025). Despite their strengths, standard GNN architectures typically assume homogeneous node and edge types, limiting their capacity to integrate and effectively leverage multi-scale data sources. Additionally, traditional GNN implementations rarely provide rigorous uncertainty quantification, which is a crucial component for supporting transparent and informed policy-making and planning decisions (Banerjee *et al.* 2003, Li *et al.* 2023b).

To address these methodological gaps, this paper proposes a Spatially Regularised Bayesian Heterogeneous Graph Neural Network (SR-BHGNN) for hierarchical small-area estimation. The SR-BHGNN approach constructs a multi-scale, heterogeneous graph structure that explicitly models both adjacency relationships (neighbourhoods) and hierarchical membership (nested census tract boundaries). Bayesian inference enables principled quantification of predictive uncertainty, enhancing the interpretability and reliability of downscaled estimates. Furthermore, inspired by Tobler's First Law of Geography (Tobler 1970, Anselin 2013), we introduce an external spatial regularisation penalty that encourages coherence in predictions between geographically proximate areas, while allowing flexibility to accommodate local heterogeneity.

We empirically validate the proposed SR-BHGNN using two real-world case studies from Greater London, UK: (1) estimating small-area population distributions and

(2) predicting local exposure to fine particulate matter (PM2.5). These cases exemplify common scenarios where fine-scale estimates are needed but detailed local data are scarce or incomplete. By systematically comparing our model's performance to established statistical benchmarks (spatial hierarchical Bayes models), conventional machine learning methods (random forests), and existing GNN architectures, we demonstrate improvements in predictive accuracy, robustness to data imbalance, and, importantly, calibrated uncertainty quantification. Additionally, we assess model performance under two distinct scenarios, complete versus limited fine-scale data availability, to illustrate its practical applicability.

To summarise, in this paper:

- We introduce SR-BHGNN, a flexible Bayesian hierarchical graph neural network explicitly designed for fine-scale estimation using coarser-scale geospatial data.
- We incorporate an external spatial regularisation term to improve robustness against spatial imbalance and enhance local predictive coherence.
- We rigorously quantify prediction uncertainties using Bayesian inference, validating these uncertainties through empirical coverage and probabilistic scoring metrics, such as the Continuous Ranked Probability Score (CRPS).
- We demonstrate the effectiveness of SR-BHGNN through two policy-relevant urban analytic tasks, fine-scale population estimation and air pollution exposure modelling, using detailed empirical studies from Greater London.

2. Background

2.1. Small-area estimation and hierarchical spatial modelling

Small-area estimation (SAE) involves the statistical prediction or imputation of attributes at fine spatial resolutions, particularly in situations where direct measurements at these scales are limited, costly, or otherwise unavailable (Rao and Molina 2015, Ghosh 2021). SAE approaches are commonly employed to address spatial mismatches between the resolution at which data are available and the finer scales at which urban planning and policy interventions must be implemented. These methods are critical for urban analytics tasks such as environmental monitoring, resource allocation, public health planning, and infrastructure development (Morales *et al.* 2021, Corral *et al.* 2022, Edochie *et al.* 2025).

In the domain of remote-sensing and raster-based spatial analysis, techniques analogous to small-area estimation, often termed 'downscaling', are well-established, and they typically involve refining coarse-resolution continuous datasets (such as satellite imagery) to produce finer-scale predictions (Wang *et al.* 2022). Techniques such as image super-resolution, data fusion, and interpolation have been extensively explored in environmental applications, including the enhancement of spatial detail in products like land surface temperature maps, precipitation maps, or vegetation indices (Atkinson *et al.* 2008, Gocht and Röder 2014, Yue *et al.* 2015, Cheng *et al.* 2024). Such methods capitalise on the spatial continuity and regularity inherent to raster data structures.

However, urban analytics predominantly involves vector-based data, structured according to administrative or political boundaries (e.g. census tracts, neighbourhoods, districts), presenting unique methodological challenges compared to raster-based contexts. Vector polygons representing these administrative units are inherently irregular, heterogeneous in size and shape, and nested within multi-level administrative hierarchies (e.g. neighbourhoods within districts, districts within cities) (Fotheringham *et al.* 2009, Fotheringham 2024). Furthermore, each polygon typically encompasses numerous attributes, including socio-economic, demographic, or infrastructural variables collected from diverse sources, such as government surveys, private-sector databases, or crowdsourced platforms (Demšar *et al.* 2013, Singleton *et al.* 2020). Consequently, standard raster-based approaches and assumptions, such as spatial continuity or uniform grid cells, do not directly transfer to the polygon-based, attribute-rich, and hierarchical context of urban spatial data (Kirilenko 2022, Mai *et al.* 2024).

Hierarchical spatial modelling methods, such as hierarchical Bayesian or empirical Bayes frameworks, have traditionally been adapted from statistics and demography for small-area estimation purposes. These models introduce multi-level random effects or partial pooling to leverage data across spatial scales and manage uncertainty in sparse local observations (Ghosh 2021, Rao and Molina 2015). Parametric Bayesian models that explicitly couple multilevel structure with horizontal spatial dependence, such as hierarchical Spatial Autoregressive (SAR) models and hierarchical Spatial Error/Spatial Moving Average (SEM/SMA) models, demonstrate how these dependencies can be jointly specified, but at the cost of strong functional forms and non-trivial identification or estimation complexity (Anselin 1988, Dong and Harris 2015, Wolf *et al.* 2021). These approaches highlight both the potential and the challenges of Bayesian formulations for capturing multi-level spatial processes, motivating the wider use of Bayesian methods. Meanwhile, existing statistical hierarchical methods typically rely on restrictive parametric assumptions, such as linear relationships or simple spatial dependence structures, which limit their effectiveness when confronting the complex non-linear interactions and heterogeneous spatial dependencies common in urban environments (Anselin 2013, Arribas-Bel and Fleischmann 2022); thus, resulting in uncertainties that are underexplored in urban analytics.

2.2. Bayesian-based urban analytics

Bayesian methods have a rich history of addressing uncertainty and hierarchical structures in spatial analysis. Early applications in disease mapping and ecological statistics demonstrated the power of hierarchical Bayesian models to capture area-level variability while borrowing strength across regions (Clayton and Kaldor 1987, Cressie 2015). In small-area estimation, hierarchical Bayes and empirical Bayes strategies improved estimates in sparsely sampled areas by introducing multi-level random effects and partial pooling (Ghosh and Rao 1994, Banerjee *et al.* 2003). These foundations have since expanded into spatial econometrics (Anselin 2013), where Bayesian formulations allow for complex spatial correlation structures and intricate dependencies among economic or demographic variables (LeSage and Pace 2009).

Recent advances in urban analytics reflect a growing interest in Bayesian modelling for large and heterogeneous datasets. For instance, Integrated Nested Laplace Approximation (INLA) has been used to efficiently approximate posterior distributions in spatially explicit models (Rue *et al.* 2009, Fuglstad *et al.* 2014, Berild *et al.* 2022), while Bayesian hierarchical frameworks have facilitated the fusion of diverse data sources (e.g. traffic flows, social media, environmental measurements) into unified predictive models (Huang and Abdel-Aty 2010, Diaconescu *et al.* 2014, Britten *et al.* 2021). These methods underscore the recognition that uncertainty quantification is essential in policy-relevant domains, where biased or imprecise estimates can lead to suboptimal decisions regarding infrastructure, healthcare, or environmental management (Atkinson and Tate 2000, Chiles and Delfiner 2012).

2.3. Graph modelling in urban studies

Graph modelling has long been pivotal in geospatial research, offering a formal representation of entities (e.g. neighbourhoods, intersections, buildings) and their interconnections (e.g. shared borders, roads, or socio-economic ties) (Ghosh *et al.* 2024). Within urban studies, these methods have been widely employed for network analysis, leveraging topological properties such as centrality, clustering, and connectivity to examine issues ranging from transportation efficiency to socio-spatial inequalities (Boeing 2017, 2022; Yap and Biljecki 2023). Recognising that contemporary urban systems often feature nested administrative structures, where local districts or neighbourhoods fall under broader jurisdictions like boroughs, counties, or metropolitan regions, recent research has explored hierarchical graph representations to account for multi-scale dependencies and varying data availability across different administrative tiers (Wang *et al.* 2021). For instance, studies on infrastructure resilience frequently model energy or water networks at both city-wide and neighbourhood-specific levels, connecting strategic nodes (e.g. power plants) with local distribution nodes (e.g. substations) in a layered graph structure (Ferrario *et al.* 2016). Similarly, small-area estimation efforts increasingly adopt multi-level graph models that ‘borrow strength’ from higher-level aggregations when local data are sparse, yet still preserve the autonomy of finer-grained units (Molina *et al.* 2014).

In recent years, Graph Neural Networks (GNNs) have gathered significant attention for modelling complex urban networks, mirroring broader shifts in spatial data science (Liu 2024). Unlike traditional raster-based methods, GNNs naturally accommodate non-Euclidean spatial structures, in which entities (e.g. neighbourhoods, intersections, or facilities) and their connections (e.g. shared boundaries, roads, or infrastructure networks) are expressed as a graph. This graph-based perspective is especially valuable for applications where relational dependencies, such as traffic flow along roads or adjacency among census tracts, strongly influence urban dynamics (Liu and Biljecki 2022). However, many GNN-oriented approaches remain confined to a single scale of analysis and produce deterministic outputs. Only recently have researchers begun to incorporate multi-level or hierarchical dimensions into GNN frameworks, aiming to capture the nested nature of geospatial boundaries while reflecting spatial dependencies across different scales (Li *et al.* 2023b, Chen *et al.* 2024, Wang and Zhu 2024, Liu *et al.* 2025a). Nevertheless, these early developments often lack rigorous uncertainty quantification and overlook the complexities of downscaling, wherein information must flow from coarser aggregates to finer units. In the following sections, we address these shortcomings by introducing a Bayesian, spatially regularised heterogeneous GNN, SR-BHGNN, that manages hierarchical urban structures, quantifies uncertainty, and facilitates downscaling for policy-relevant applications.

3. Method

Figure 1 illustrates the proposed SR-BHGNN framework, which comprises two main components: hierarchical spatial graph construction and the SR-BHGNN model. In the following sections, we introduce each component, detailing its respective structures and roles within the overall architecture. It is worth noting that in this paper, we use the term 'scale' to refer specifically to the horizontal grouping/adjacency structure at a given geospatial unit level (i.e. OAs and LSOAs, see Figure 1 and Section 4). We distinguish this structural notion of scale from 'ancillary data', which in the small-area estimation literature typically refers to independent, externally obtained datasets that are combined with core data through a separate modelling step (e.g. dasymetric weights, remote-sensing covariates) (Rao and Molina 2015, Dong and Harris 2015). In our experiments, the only information from higher levels used by the hierarchical models is the grouping structure and the variables already measured at those levels; we do not introduce external ancillary data sources.

3.1. Hierarchical spatial graph

To construct a multi-level, heterogeneous, hierarchical graph, we draw upon two sets of statistical geography units (i.e. census tracts, see Section 4), corresponding to different spatial scales: larger spatial units and smaller spatial units. Each unit is associated with a set of features (e.g. socio-economic indicators, environmental measurements,

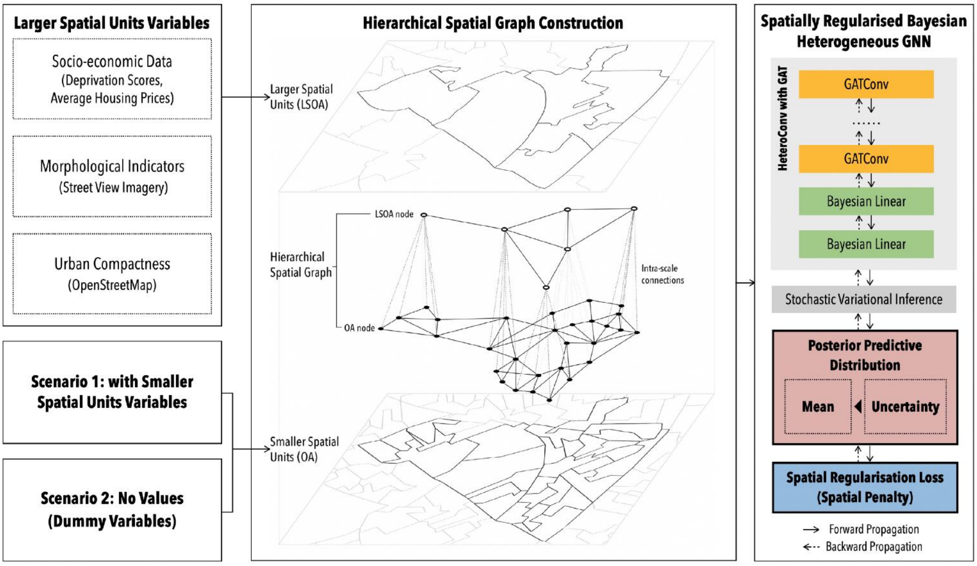

Larger Spatial Units Variables
Socio-economic Data
(Deprivation Scores,
Average Housing Prices)
Morphological Indicators
(Street View Imagery)
Urban Compactness
(OpenStreetMap)
Scenario 1: with Smaller Spatial Units Variables
Scenario 2: No Values (Dummy Variables)
Hierarchical Spatial Graph Construction
Larger Spatial Units (LSOA)
LSOA node
Hierarchical Spatial Graph
Intra-scale connections
OA node
Smaller Spatial Units (OA)
Spatially Regularised Bayesian Heterogeneous GNN
HeteroConv with GAT
GATConv
......
GATConv
Bayesian Linear
Bayesian Linear
Stochastic Variational Inference
Posterior Predictive Distribution
Mean
Uncertainty
Spatial Regularisation Loss (Spatial Penalty)
Forward Propagation
Backward Propagation

Figure 1. The framework for the proposed SR-BHGNN.

Table 1. List of OA- and LSOA-level variables used in the study.

Spatial level	Dataset	Variables	Mean	Standard deviation
LSOA	England IoD	Barriers to Housing and Services	31.43	9.80
		Living Environment	28.88	11.03
		Multiple Deprivation	21.28	10.93
	Google Street View	Sidewalk	0.35	0.24
		Building	0.83	0.50
		Road	0.97	0.89
		Terrain	0.11	0.22
		Greenery	0.59	0.50
	OpenStreetMap	Road Density	0.77	0.39
		Building Density	0.58	0.31
OA	LOAC variables	Ownership or shared ownership	0.82	0.12
		Social rented	0.58	0.27
		Private rented	0.73	0.13
		Occupancy rating of rooms: +1 or more	0.73	0.09
		Occupancy rating of rooms: −1 or less	0.51	0.18
		Standardised Disability Ratio	0.78	0.05
		Provides no unpaid care	0.77	0.08
		2 or more cars or vans in household	0.62	0.20
		Highest level of qualification: Level 1, 2 or Apprenticeship	0.56	0.14
		Highest level of qualification: Level 3 qualifications	0.52	0.08
		Highest level of qualification: Level 4 qualifications or above	0.69	0.14
		Job Type: Part-time	0.77	0.06
		Job Type: Full-time	0.84	0.05

Mean and standard deviation are reported to indicate the central tendency and dispersion of each covariate prior to modelling, providing context for the scale of the input features.

see details in Section 4) and, for smaller units, a target variable to be predicted. The final graph encodes horizontal adjacency within each scale and vertical membership relationships between scales, forming the foundation for the Bayesian GNN approach discussed in Section 3.2. An illustrative figure can be found in the left part of Figure 1.

Within each spatial scale, spatial adjacency is determined using a Queen contiguity (Rey *et al.* 2022), whereby two polygons are deemed neighbours if they share at least one boundary point or vertex; these edges facilitate horizontal message passing within each scale, capturing local spatial dependencies. Queen contiguity is a standard and widely adopted specification for irregular spatial units (e.g. census tracts), commonly used in spatially explicit geospatial artificial intelligence (GeoAI) applications for urban modelling (De Sabbata and Liu 2023, Liu *et al.* 2025a). Nonetheless, alternative adjacency definitions, such as Rook contiguity, barrier-aware contiguity, or centroid-based k-nearest-neighbour graphs, are equally feasible. We report the results for these variants in Table A1 of Appendix A to demonstrate the robustness of our findings to the choice of spatial graph.

Membership edges are established between smaller and larger units based on spatial containment, reflecting the hierarchical nature of urban systems. If a smaller unit's geometry lies entirely within a larger unit's boundary, an undirected edge is created between the smaller and the larger. Formally, for smaller unit s and larger unit ℓ:

$$s \in \ell \Rightarrow (s \rightarrow \ell) \; \mathit{and} \; (\ell \rightarrow s) \tag{1}$$

Therefore, these cross-level connections enable vertical information transfer, allowing the GNN to model multi-scale interactions.

By combining node features, target labels (categorical labels or numerical values) for the smaller units, adjacency edges at each scale, and membership edges across scales, we obtain a heterogeneous graph structure suitable for hierarchical analysis. Formally, the graph contains:

- **Node sets:**
  1. $\mathbf{X}_{\text{large}} \in \mathbb{R}^{N_{\text{large}} \times d_\ell}$: feature matrix for larger spatial units;
  2. $\mathbf{X}_{\text{small}} \in \mathbb{R}^{N_{\text{small}} \times d_s}$: feature matrix for smaller spatial units;
  3. $\mathbf{y}_{\text{small}} \in \mathbb{Z}^{N_{\text{small}}}$: target labels for smaller spatial units.
- **Edge sets:**
  1. Spatial adjacency among larger units (horizontal dependency);
  2. Spatial adjacency among smaller units (horizontal dependency);
  3. Membership edges linking smaller to larger units and vice versa (vertical dependency).

The membership links are handled as one relation type with two directions; they allow information to flow ‘upwards’ and ‘downwards’ between scales but share the same learnable parameters for this relation. As such, this two-tiered graph captures the nested hierarchy of census tracts and local spatial dependencies, enabling a structured representation of urban spatial systems. It forms the input to the proposed SR-BHGNN described in Section 3.2.

3.2. Spatially regularised Bayesian heterogeneous graph neural network

The proposed framework operates on the above-mentioned *heterogeneous hierarchical graph* $\mathcal{G} = (\mathcal{V}_L \cup \mathcal{V}_S, \mathcal{E})$, in which $\mathcal{V}_L$ denotes the set of *larger* census tract units and $\mathcal{V}_S$ denotes the set of *smaller* census tract units. Each node $v \in \mathcal{V}_L \cup \mathcal{V}_S$ may possess a feature vector x_v, with dimension possibly differing by node type. In particular, two configurations are considered for the smaller units: one in which each smaller unit carries its vector of attributes and another in which the feature matrix for smaller units is effectively empty, thereby relying on adjacency- and membership-based message passing for representation learning. Adjacency edges in $\mathcal{E}$ capture spatial contiguity within each census tract scale, while membership edges link smaller units to the larger units containing them and vice versa.

To handle this multi-scale structure, the model employs a *heterogeneous graph neural network* (Fey and Lenssen 2019), wherein each edge type (adjacency or membership) has its own message-passing transformations. Specifically, let $h_v^{(l)} = x_v$ be the initial features of node v. At layer $l+1$, the updated embedding $h_v^{(l+1)}$ is computed by aggregating messages from the neighbour set $\mathcal{N}(v)$, which may include nodes of the same type (adjacent) or different types (membership). For each edge relation r, a trainable weight matrix or attention mechanism transforms the incoming messages. Denoting $\bigoplus$ as an aggregation operator (e.g. mean aggregator in this paper) and $\sigma(\cdot)$ as a non-linear activation:

$$h_v^{(l+1)} = \sigma\left(W_r^{(l)}\left[h_v^{(l)} \parallel \bigoplus_{u \in \mathcal{N}_r(v)} h_u^{(l)}\right] + b^{(l)}\right) \tag{2}$$

where $\mathcal{N}_r(v)$ is the set of neighbours of v under relation r, and $\|$ denotes a concatenation of embeddings. Because each edge type (within-OA, within-LSOA, and membership links) is parametrised separately, SR-BHGNN does not force a single neighbourhood size across the whole graph. Instead, the model can learn one pattern of message passing within a level and another for cross-level connections. In practice, this means it can decide how strongly to propagate information among nearby units within the same scale and how much to transfer between scales.

In addition to being heterogeneous, the model is *Bayesian*. Each weight matrix $W_r^{(l)}$ and bias $b^{(l)}$ is drawn from a prior distribution, often assumed Gaussian, of the form:

$$W_r^{(l)} \sim \mathcal{N}(0, \alpha^2 I), \quad b^{(l)} \sim \mathcal{N}(0, \beta^2 I) \tag{3}$$

where $\alpha^2, \beta^2 > 0$ are fixed hyperparameters. Rather than treating these parameters as point estimates, variational inference approximates their posterior distributions, thus capturing parameter uncertainty. We define a variational distribution $q_\phi(\theta)$ over all model parameters θ, typically taken as a fully factorised Gaussian:

$$q_\phi(\theta) = \prod_{i=1}^{|\theta|} \mathcal{N}(\theta_i; \mu_i, \sigma_i^2) \tag{4}$$

where $\phi = \{\mu_i, \sigma_i\}$ are the learnable variational parameters, which allows for scalable Bayesian inference using the reparameterisation trick during training (Kingma *et al.* 2015). Unlike full posterior sampling techniques such as Markov Chain Monte Carlo (MCMC), which are computationally prohibitive for deep neural models on large spatial graphs, variational inference offers a scalable, optimisation-based framework suitable for large spatial graphs with complex architectures (Blundell *et al.* 2015, Blei *et al.* 2017, Zhang 2019).

After several layers of message passing, the model produces an output vector z_v for each node $v \in \mathcal{V}_S$, which can be interpreted according to the nature of the prediction task. In classification settings, a softmax function is applied to obtain predicted class probabilities:

$$P_v = \text{softmax}(z_v) \tag{5}$$

and the likelihood of the observed label y_v is given by a categorical distribution. In regression tasks, by contrast, the output z_v can represent the mean of a continuous distribution, such as a Gaussian. Denoting all model parameters by θ, the likelihood of the observed outcome y_v is defined as:

$$p(y_v|z_v, \theta) = \begin{cases} \text{Cat}(\text{softmax}(z_v)), & \text{classification}, \\ \mathcal{N}(y_v|z_v, \sigma^2), & \text{regression}, \end{cases} \tag{6}$$

where σ^2 may be fixed or learned as a function of θ. In this way, the model provides a unified Bayesian framework for discrete and continuous prediction tasks on hierarchical spatial graphs.

To promote smoothness in the predictions among spatially adjacent smaller units, the model includes a *spatial regularisation* term that penalises large discrepancies in predicted target probabilities. Concretely, if P_v is the predicted target probability vector for node $v \in \mathcal{V}_S$, and $\mathcal{E}_S \subseteq \mathcal{E}$ is the set of adjacency edges among the smaller units, then the regularisation term is defined by

$$\Omega_{\text{spatial}}(\theta) = \lambda \sum_{(v,w)\in\mathcal{E}_S} \| P_v - P_w\|^2 \tag{7}$$

where $\lambda > 0$ is a regularisation weight and a hyperparameter in the model. Note that such a regularisation term is external to the Bayesian posterior and does not constitute a prior over parameters. It directly penalises differences in model outputs across neighbouring areas to encourage spatial coherence.

The total objective function thus merges the log-likelihood for all observed labels (under the Bayesian parameter priors) with the penalty on spatial inconsistency defined above. Specifically, again, let $q_\phi(\theta)$ be the variational distribution over model parameters θ, and let $p(\theta)$ be the prior. Then, the negative evidence lower bound (ELBO) is

$$\mathcal{L}_{\text{ELBO}}(q_\phi) = \mathbb{E}_{q_\phi(\theta)}[-\log p(\{y_v\}|\theta)] \ + \ \text{KL}\big(q_\phi(\theta) \parallel p(\theta)\big) \tag{8}$$

where $p(\{y_v\}|\theta)$ is the product of the individual likelihoods for all observed y_v. KL denotes the Kullback–Leibler divergence (Kullback and Leibler 1951). Accordingly, the overall training loss is given by

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}}(q_\phi) \ + \ \Omega_{\text{spatial}}(\theta) \tag{9}$$

During training, a stochastic variational inference procedure estimates $q_\phi(\theta)$ by minimising $\mathcal{L}$, thus balancing good predictive performance with uncertainty quantification and spatial consistency.

At test time, the approximate posterior over model parameters can be sampled multiple times to generate a set of output vectors $\{z_v^{(s)}\}_{s=1}^S$ for each node $v \in \mathcal{V}_S$. These outputs are then used to compute predictive summaries. In classification settings, softmax is applied to each sample to obtain class probabilities, and the final predictive distribution is obtained by averaging:

$$\hat{P}_v = \frac{1}{S}\sum_{s=1}^{S} \text{softmax}(z_v^{(s)}), \tag{10}$$

with the sample variance across $\{z_v^{(s)}\}$ or $\{\hat{P}_v^{(s)}\}$ serving as a measure of predictive uncertainty. In regression settings, the mean prediction is taken as the average of the sampled outputs,

$$\hat{y}_v = \frac{1}{S}\sum_{s=1}^{S} z_v^{(s)}, \tag{11}$$

while the sample variance provides an estimate of uncertainty in the predicted value. Such a capacity to quantify uncertainty is particularly beneficial in small-area estimation contexts, where observations at finer spatial resolutions may be sparse, noisy, or altogether unavailable.

Section 4 illustrates the model's practical utility by applying it to two empirical examples, highlighting its flexibility and robustness in real-world urban analytics.

3.3. Model implementation

The proposed SR-BHGNN is implemented in Python using the PyTorch (Ansel *et al.* 2024) and PyTorch Geometric (Fey and Lenssen 2019) libraries. The heterogeneous graph data structures, along with adjacency and membership edges, are constructed using PyTorch Geometric's built-in mechanisms for managing multi-type node and edge relations. To facilitate Bayesian inference, the implementation employs Pyro (Bingham *et al.* 2019), which provides a flexible interface for defining priors, variational guides, and stochastic variational inference routines.

Regarding hyperparameters, our experiments suggest that a heterogeneous GAT-based message-passing block with one convolutional kernel per relation type (see Section 3.1) offers a good balance between model capacity and computational efficiency. This constitutes one message-passing layer in the sense of GNN depth, but with distinct weights learned for each edge type. The outputs are then passed through Bayesian linear layers with ReLU activations to produce hidden representations, and finally, a Bayesian linear output layer for OA-level predictions. In our experiments, the hidden dimension is set to 16, which provides a good compromise between capacity and speed, although this choice may be adjusted as data complexity varies. We typically choose a learning rate of 0.001 with *ClippedAdam* as the optimiser to stabilise training when sampling from posterior distributions. The strength of the spatial regularisation term λ is tuned by comparing validation performance on a small grid of values (e.g. $\{0.01, 0.05, 0.1, 0.5, 0.9\}$; empirical results indicate that a moderate level of spatial smoothing of $\lambda = 0.1$ best controls local outliers without eroding fine-grained patterns for the regression task and $\lambda = 0.5$ for the classification task. For continuous outcomes (e.g. population density), as in regression tasks, we adopt a normal likelihood with a Bayesian linear output layer. Each forward pass samples parameters from the approximate posterior, enabling robust uncertainty quantification over predicted means. All training protocols use the Evidence Lower Bound (ELBO) as the objective function, with model convergence monitored by the evolution of ELBO and accuracy metrics on a held-out validation set, and early stopping applied with a patience of 200 epochs. To ensure full reproducibility, a fixed random seed (42) is set for all experiments and documented in the accompanying code repository.

4. Empirical analysis

This section demonstrates the proposed SR-BHGNN on two real-world tasks within a common urban setting. Section 4.1 outlines the data processing pipeline for Greater London to assemble multi-level spatial units, socio-economic indicators, and physical environment descriptors. Section 4.2 presents our approach to population estimation at finer spatial scales, comparing predictive performance with four baselines and under different data availability scenarios, demonstrating the effectiveness of the proposed spatial regularisation term in the model. Section 4.3 then extends the framework to air pollution prediction, illustrating how the model handles the predictions of environmental variables and highlighting the practical usefulness of uncertainty quantification and its derived insights. Through these two complementary examples, we showcase the flexibility of the method in addressing diverse small-area estimation and forecasting challenges.

4.1. Data processing

This study focuses on Greater London as an empirical testbed, owing to its well-established census tract hierarchy and the availability of diverse socio-economic and spatial datasets. As demonstrated in Figure 1, we adopt two nested census tract levels: LSOAs and OAs, with each OA spatially contained within a single LSOA. Such a nested arrangement is conducive to evaluating the proposed hierarchical modelling framework, as it allows information flow between coarser (LSOA) and finer (OA) scales. The LSOAs and OAs used in this research are based on UK Census 2021 geographies (Office for National Statistics 2021).

The primary target variables at the OA scale are total population density and annual average PM $_{2.5}$ concentrations. The OA-level population data is obtained from the 2021 UK Census provided by Office for National Statistics (2021). Although OA populations are bounded by design, each area must contain a minimum of 100 residents, with a target average of approximately 300; substantial variation still occurs, particularly in diverse urban environments such as London. Despite such design constraints, population density aggregated at OA levels remains a critical modelling target, as they directly inform resource allocation, service delivery, and infrastructure planning at the neighbourhood level (Singleton and Longley 2024, Wyszomierski *et al.* 2024). In the meantime, the raw data comprise PM $_{2.5}$ concentrations and exposure values from 2013 (Greater London Authority 2017b), originally aligned with the 2011 OA boundaries. Since our overall framework relies on the 2021 UK census geographies, we first *rebased* the pollutant dataset to the newer OA geographies by overlaying maps and interpolating missing or mismatched values using an official lookup table. Where direct matches were unavailable, we performed qualitative checks on boundary overlaps and used domain expertise to reconcile minor discrepancies; thus, we ensured consistency across all spatial layers. Both values (population and PM $_{2.5}$) represent the ‘ground truth’ for model evaluation and are used for both training and testing purposes. In the model testing, we simulate a realistic use case where OA-level target variables are only available for a subset of areas. In each task, the Output Areas were randomly partitioned into 70% training and 30% held-out test sets, a protocol widely used in GeoAI applications (Zhu *et al.* 2020, Wang and Zhu 2024, Liu *et al.* 2025a, 2025b); a spatial block cross-validation is presented in Appendix B to assess robustness to spatial leakage. Our model and the baseline models (see Section 4.1) are trained exclusively on the training OAs and evaluated on the held-out test set. Such a set-up mimics scenarios where fine-grained target variables are known for some small areas (e.g. pilot surveys, selectively monitored locations) and need to be estimated elsewhere based on aggregated data and structural priors.

A summary of the input data used at each level is presented in Table 1. All variables described in the table are used consistently in both case studies; only the outcome variable differs (population in Case Study 1 and PM $_{2.5}$ in Case Study 2, see subsections below). At the LSOA level, we compile several node attributes to capture both the socio-economic and physical dimensions of urban life. First, *deprivation scores* (*Barriers to Housing and Services*, *Living Environment*, and *Multiple Deprivation* from the England Index of Multiple Deprivation (Trust for London 2019) (rebased to 2021 UK census geographies) and average housing prices between 1995 and 2017 (Greater London Authority 2017a) are included to reflect spatial variations in socio-economic well-being. These measures are vital for population and environmental modelling as highly deprived areas often exhibit different demographic profiles, health outcomes and housing conditions.

Second, we incorporate urban morphological indicators derived from 230,971,036 *Google Street View* (GSV) images, processed through semantic segmentation using the Mask2Former model (Cheng *et al.* 2022) pre-trained on the Mapillary Vistas Dataset (Neuhold *et al.* 2017, Ito *et al.* 2025). All imagery was accessed via the Google Street View Static API under the research Terms of Service (Google LLC 2025) as of 13 February 2025. For each street segment, we utilised all available images dated 2019–2025 within London; no additional subsampling was applied beyond the availability of the API. GSV imagery provides ground-level perspectives that capture a granular snapshot of the built environment, including subtle details and variations often overlooked by remote sensing or aggregated data (Biljecki and Ito 2021, Fan *et al.* 2023). By systematically identifying features such as building façades, terrain, greenery, sidewalks, and roads, these derived data (i.e. proportion of pixels based on the image segmentation results) which were mean-pooled to the census tracts used in this study and enrich the model's understanding of how local urban form correlates with both population distributions and air pollution patterns, reflecting everyday physical conditions in a way that overhead imagery (e.g. remote sensing) may not.

Finally, we measure *road and building density* using OpenStreetMap (OSM) since transport infrastructure and urban compactness influence human mobility, exposure levels, and resource allocation (Shi *et al.* 2016, Li *et al.* 2023a, Cao and Su 2024). OSM features were extracted as of 3 March 2025. For London, OSM provides reasonable coverage for the structural indicators used here (e.g. road length/density, built-form proxies) (Biljecki *et al.* 2023), and we therefore use it as is. Nevertheless, as OSM is a crowd-sourced dataset whose completeness and positional accuracy can vary by feature type and location, we acknowledge any residual omissions or inaccuracies as a potential source of uncertainty and a limitation of our analysis. The target variables (population and PM $_{2.5}$) and predictor variables are aligned as closely as possible in time. Because the built-form and street-network indicators used here are relatively stable over time (Clifton *et al.* 2008, Barrington-Leigh and Millard-Ball 2020, Tümtürk *et al.* 2024), we expect limited impact from minor temporal mismatches, which we explicitly note as a limitation of the analysis (see Discussion). Yet, a comprehensive audit of completeness and temporal alignment is beyond the scope of this paper.

For the OA level, as shown in Figure 1, we examine two scenarios reflecting different data availability conditions. In the first scenario, we exploit *census-derived variables* (e.g. proportions of owner-occupied or socially rented dwellings) and socio-economic indicators (health and education metrics) collected from the London Output Area Classification (LOAC, a geodemographic that summarises the built and population characteristics in London) (Singleton and Longley 2024) and 2021 UK Census statistics

(Office for National Statistics 2021). These finer-grained attributes offer additional insight into local demographic and service-related factors that influence population estimates and air pollution outcomes. In the second scenario, OAs have *no node-specific features* beyond their adjacency and membership connections, allowing the model to rely solely on structural relationships and LSOA-level signals to infer local characteristics. This approach assesses how effectively the hierarchical GNN can compensate for missing fine-scale covariates. Testing the model under both scenarios quantifies how direct OA features enhance predictive performance or whether structural embedding and Bayesian uncertainty estimates suffice when smaller-unit data are unavailable.

It is worth noting that, given the broad array of socio-economic, morphological, and infrastructural variables being integrated, some degree of correlation among these features is inevitable. For instance, areas scoring high on certain deprivation measures may also exhibit distinctive GSV-based characteristics (e.g. fewer green spaces or denser building façades). Rather than eliminate overlapping predictors, we retain them to exploit the full representational capacity of the model and to reflect the complexity of real-world environments. In particular, the hierarchical GNN's ability to learn non-linear relationships and partial dependencies often reduces the risk of adverse effects from multicollinearity (Vatcheva *et al.* 2016, Gao *et al.* 2023). Meanwhile, the Bayesian inference framework provides posterior estimates that can reveal how these correlated features jointly contribute to the model's uncertainty and predictions.

These LSOA- and OA-level inputs compose a multi-scale dataset reflecting socio-economic, morphological, and infrastructural characteristics across Greater London. After integrating the data into the LSOAs and OAs, Queen contiguity is applied to construct spatial graphs from the polygon data within their scales. By integrating multiple data sources at different granularities, this design enables a robust evaluation of the proposed framework under realistic conditions of partial or missing covariates at finer scales. Subsequent sections detail how the hierarchical graph is used to facilitate *population estimation* and *air pollution prediction*, illustrating the model's versatility in addressing diverse urban analytics tasks.

4.2. Case study 1: population estimation

Accurate population estimates at fine spatial scales are essential for a variety of urban policy and planning tasks, including service provision, infrastructure development, and emergency resource allocation (Sun 1971, McDonald 1989, Gottdiener *et al.* 2015, Zoraghein and Leyk 2018, Sinha *et al.* 2019, Schnake-Mahl *et al.* 2020, Singleton *et al.* 2020). However, conventional census data tend to be aggregated at coarser census tract levels and updated infrequently, thereby limiting their value for real-time or localised decision-making. To address these constraints, we apply the proposed SR-BHGNN to estimate the population at the Output Area (OA) level, illustrating its practical relevance and methodological advantages when smaller-scale data are sparse or inconsistent. Moreover, the OA population distribution in Greater London, similar to many other urban settings (Newbold 2021), is markedly imbalanced (skewness $\approx$ 47.01, kurtosis $\approx$ 4603.82), such that most OAs have modest resident density while a small subset hosts disproportionately large populations. This extreme imbalance introduces considerable challenges for predictive modelling. In the following analysis, we therefore evaluate how effectively the SR-BHGNN, alongside several baselines, handles this demanding yet realistic scenario of small-area population estimation.

We examine two formulations of the population task, classification and regression, under two OA feature settings: with OA-level covariates (SR-BHGNN_OACensus) and without them (SR-BHGNN_OADummy). For classification, continuous population values are binned into three categories via Jenks' natural breaks (Jenks 1967); regression directly predicts continuous densities.

We select baselines that (i) represent canonical families familiar to the small-area estimation community, (ii) are widely used and reproducible with standard implementations, (iii) operate on lattice/census tract units without bespoke engineering, (iv) support both continuous and categorical targets (or have straightforward adaptations), and (v) scale to city-wide experiments. Accordingly, we benchmark SR-BHGNN against five baselines and include one ablation:

- Random Forest (RF). A strong, non-parametric baseline repeatedly shown to perform well in spatial interpolation and downscaling while remaining implementation-agnostic and robust (Sekulić *et al.* 2020, Maxwell *et al.* 2021). Its popularity and reproducibility make it an appropriate off-the-shelf comparator. We use the default implementation in Pedregosa *et al.* (2011) with 500 trees and no feature selection, trained and tested on the same OA-level observations as the other models.
- Geographical Random Forest (GRF). We include a geographically weighted RF (Georganos and Kalogirou 2022) as a spatially adaptive, non-hierarchical baseline. For each prediction location, a local random forest is estimated using the training data, with sample weights determined by a Gaussian distance–decay kernel. The bandwidth is selected by cross-validation on the training set over a grid tied to the empirical distance distribution. For computational efficiency, each local fit is restricted to the k-nearest training units (KD-tree; $k = 200$). All other hyperparameters mirror the global RF baseline (500 trees). This procedure yields spatially localised, non-parametric predictions for both regression and classification tasks.
- Graph Attention Network (GAT). A state-of-the-art single-scale graph neural baseline (Velickovic *et al.* 2017). Because our contribution in this paper advances both interpolation and hierarchical GNN design, GAT isolates the added value of explicit hierarchy and Bayesian inference, rendering it an ideal baseline comparison.
- GAT-Concat (naïve multi-scale). To ensure a fair comparison, we include a 'GAT-Concat' baseline that injects LSOA-level features into each OA node via simple concatenation, allowing this model to access the same multi-level information as SR-BHGNN without explicitly modelling the hierarchical structure; thus, testing whether feature fusion alone is sufficient. For each OA within an LSOA, we allocate LSOA-level features proportionally by OA area share and concatenate them with OA features – thus incorporating multi-scale information without explicit membership edges. In Appendix A, we additionally report two analogous baselines, RF-Concat and GRF-Concat, which extend the RF and GRF models with the same

concatenated OA-LSOA feature set. These comparisons assess whether simple data downscaling, without hierarchical message passing, can aid in prediction tasks.

- Spatial Hierarchical Bayes (Spatial HB). A well-established workhorse for small-area estimation that pools information via hierarchical structure and encodes lattice dependence with conditional autoregressive (CAR) priors (Rao and Molina 2015, Whitworth *et al.* 2017, Wall 2004). Following Wikle *et al.* (1998), we implement a Bayesian small-area model that transfers information from LSOAs to OAs, with a CAR prior over adjacent OAs and LSOA-level features as group-level predictors. Inference uses Gibbs sampling (10,000 iterations; 2,000 burn-in) (Geyer 1992), based on the same OA covariates as RF and GAT. For classification, we adapt a hierarchical ordinal probit structure (Rampichini and Schifini d'Andrea 1998) to enable probabilistic categorisation.
- BHGNN (ablation). A heterogeneous GNN with Bayesian parameter inference without the spatial regularisation term. Allowing OA-level features, this ablation isolates the specific effect of the spatial smoothness constraint relative to SR-BHGNN.

In Appendix A, we further compared the model with two additional models (a Besag–York–Mollié model and a Graph Laplacian–regularised ridge), which are commonly used in small-area estimations. Although our comparative evaluation focuses on baselines introduced above, we acknowledge that other established approaches, such as spatial regression (Chi and Zhu 2008), dasymetric modelling (Briggs *et al.* 2007), and kriging/Gaussian process (GP) interpolation (Oliver and Webster 1990, Bajjali 2023), are widely used and have advanced small-area estimation. While these approaches offer well-documented advantages, they are typically predicated on specific assumptions, such as linear predictor–response relationships or stationary covariance structures defined on continuous supports, and in many cases require ancillary data sources (e.g. high-resolution land-use information or dense monitoring networks) (Risser and Calder 2015, Tong *et al.* 2022). In highly heterogeneous urban contexts such as Greater London, these assumptions and data dependencies can constrain their applicability to categorical outcomes on irregular census tract units and to explicitly hierarchical problem settings. For this reason, in the present study, we adopt the Spatial HB model as the canonical hierarchical–lattice comparator for small-area estimation, rather than as a proxy for kriging/GP or SAR/SEM families, which operate on different spatial supports and embody distinct dependence structures. The proposed SR-BHGNN is intended to complement this tradition by relaxing linearity and stationarity assumptions and by jointly modelling heterogeneous features and hierarchical spatial dependencies within a single, uncertainty-aware framework.

Table 2 presents the comparative results. In the classification task, *SR-BHGNN_OACensus* achieves the highest Accuracy (0.85) and F-score (0.81), supported by strong Recall (0.72) and Precision (0.92). Even without access to detailed OA-level features, *SR-BHGNN_OADummy* remains highly competitive (Accuracy = 0.74, F-score = 0.62), illustrating the importance of spatial regularisation and hierarchical reasoning even when fine-grained covariates are absent.

In contrast, RF and GAT baselines, which operate solely on flat OA-level inputs or simple adjacency structures, yield substantially lower Accuracy ($\leq$0.66) and F-score

Table 2. Performance comparison of population estimation methods across classification and regression tasks.

Metric	SR-BHGNN_ OACensus	SR-BHGNN_ OADummy	BHGNN	RF	GRF	GAT	GAT-Concat	Spatial HB
Classification								
Accuracy	0.85	0.74	0.76	0.61	0.66	0.63	0.33	0.58
Recall	0.72	0.62	0.33	0.55	0.57	0.58	0.27	0.58
Precision	0.92	0.61	0.25	0.60	0.59	0.59	0.29	0.57
F-score	0.81	0.62	0.28	0.57	0.57	0.58	0.30	0.57
Regression								
R^2	0.58	0.41	0.02	0.12	0.16	0.18	0.09	0.33
RMSE	7234.19	10083.26	22378.90	14681.24	14892.01	14319.02	19076.11	12160.81
MAPE	39.78%	45.28%	163.47%	107.26%	105.26%	101.16%	124.81%	51.07%

SR-BHGNN refers to the proposed Spatially Regularised Bayesian Heterogeneous Graph Neural Network with either OA census features (OACensus) or dummy features (OADummy).

($\leq$0.58). GAT-Concat, which naively combines multi-scale features through simple concatenation without modelling their structural dependencies, performs markedly worse (Accuracy = 0.33, F-score = 0.30). The similarly poor results of RF-Concat and GRF-Concat, as shown in Appendix Table A1 (RF-Concat: Accuracy = 0.49, F-score = 0.46; GRF-Concat: Accuracy = 0.39, F-score = 0.33), further reinforce that merely injecting additional higher-level features, without representing their hierarchical relationships, tends to degrade predictive performance by introducing noise and obscuring the distinction between local and contextual patterns. Spatial HB achieves moderate classification performance (Accuracy = 0.58, F-score = 0.57), confirming the value of cross-scale information pooling. Yet, it lacks the flexibility and adaptability enabled by learned, non-linear hierarchical embeddings in SR-BHGNN. Finally, BHGNN, without explicit spatial regularisation, underperforms across all metrics, with an Accuracy of 0.76 but critically low Precision (0.25), Recall (0.33), and F-score (0.28), suggesting it struggles to resolve the imbalanced and spatially imbalanced population distributions present in the data.

For the regression task, *SR-BHGNN_OACensus* again leads, achieving an R^2 of 0.58. Even using only dummy OA features, *SR-BHGNN_OADummy* secures an R^2 of 0.41, outperforming Random Forest ($R^2 = 0.12$), GRF ($R^2 = 0.16$), and GAT ($R^2 = 0.18$), both with RMSEs exceeding 14,000. GAT-Concat again demonstrates poor performance, confirming that naive feature fusion without hierarchical structuring fails to capture the complexity of spatial population patterns. Although spatial HB captures some cross-scale structure ($R^2 = 0.33$, MAPE = 51.07%), it is ultimately less expressive and less accurate than the graph-driven, Bayesian approach. BHGNN deteriorates significantly in regression ($R^2 = 0.02$, RMSE = 22378.90, MAPE = 163.47%), reinforcing the need for spatial smoothness and hierarchical integration to model highly imbalanced, spatially heterogeneous population data effectively.

To further assess the spatial validity of the population estimates, we analysed the spatial autocorrelation of model residuals using Moran's I statistic (Moran 1950), as illustrated in Figure 2. Clear distinctions emerge among the models regarding the spatial distribution of their residuals. The SR-BHGNN demonstrates minimal spatial clustering (Moran's $I = 0.08$, $p < 0.01$), indicating a more spatially random pattern of prediction errors. In sharp contrast, the Random Forest residuals exhibit substantial spatial autocorrelation (Moran's $I = 0.78$), suggesting systematic over-predictions (high-

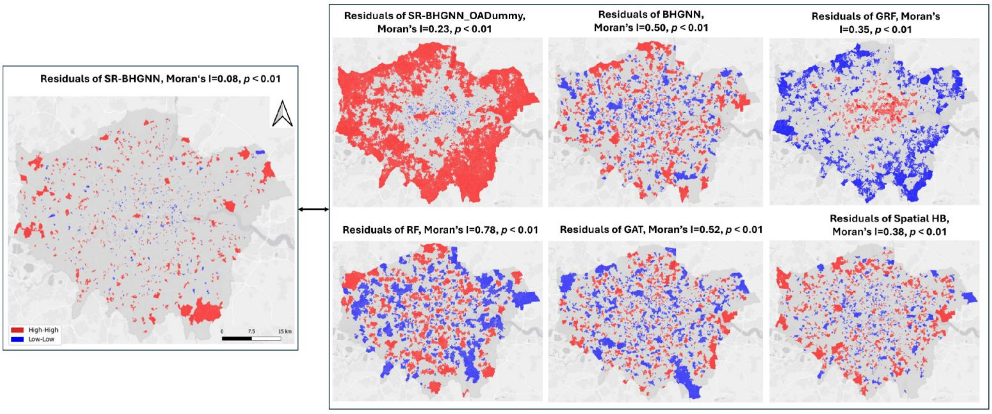
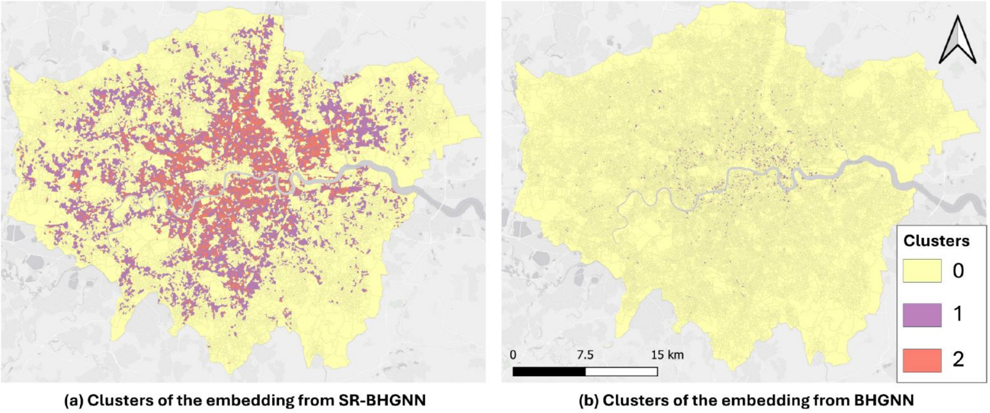

Residuals of SR-BHGNN, Moran's I=0.08, $p < 0.01$
🟥 High-High
🟦 Low-Low
0 7.5 15 km
Residuals of SR-BHGNN_OADummy, Moran's I=0.23, $p < 0.01$
Residuals of BHGNN, Moran's I=0.50, $p < 0.01$
Residuals of GRF, Moran's I=0.35, $p < 0.01$
Residuals of RF, Moran's I=0.78, $p < 0.01$
Residuals of GAT, Moran's I=0.52, $p < 0.01$
Residuals of Spatial HB, Moran's I=0.38, $p < 0.01$

Figure 2. Spatial distribution of the residuals of each model produced in the regression task. Cross-model visual comparison (SR-BHGNN *vs* baselines) of the distribution patterns of residuals, with red areas indicating over-prediction and blue areas indicating under-prediction. Moran's I values (reported in Section 4.2) quantify the global spatial autocorrelation of residuals. At the same time, the LISA maps highlight statistically significant local clusters: high–high (red) and low–low (blue) clusters indicate spatially correlated over- and under-predictions, respectively, whereas grey areas denote non-significant local associations. All maps share the same legend, north arrow, and scale bar for ease of comparison.

high clusters) across outer boroughs and under-predictions (low-low clusters) concentrated within inner boroughs. Intermediate levels of spatial autocorrelation are observed for GAT (0.52), BHGNN (0.50), GRF (0.35), and Spatial HB (0.38), each showing noticeable residual clusters. Notably, the SR-BHGNN_OADummy variant, which excludes detailed OA-level features, exhibits increased spatial autocorrelation (Moran's $I = 0.23$), underscoring the importance of incorporating detailed, multi-scale information. These results reinforce the advantage of employing spatial regularisation and hierarchical message passing in capturing complex urban structures, ensuring more robust and unbiased population estimations in metropolitan contexts such as London.

Although classification and regression metrics confirm that the proposed SR-BHGNN outperforms its non-regularised counterpart (BHGNN) in handling highly imbalanced population data, it is equally illuminating to visualise and analyse the node embeddings learned by each model. Specifically, we extract the final hidden-layer representations of all OAs from both SR-BHGNN and BHGNN, then cluster those embeddings via the K-Means clustering method ($K = 3$) and map each OA's cluster membership. As shown in Figure 3(a), the clustering derived from the SR-BHGNN embeddings exhibits a pronounced spatial grouping: OAs belonging to the same cluster tend to form continuous patches rather than appear randomly scattered. This pattern suggests that the model's spatial regularisation effectively encodes adjacency constraints in the latent space, encouraging physically contiguous zones, particularly those with similar demographic or environmental characteristics, to have embeddings that lie closer together. By contrast, the BHGNN embeddings shown in Figure 3(b) produce clusters that appear largely dispersed over Greater London, echoing the weaker performance observed in the model's classification and regression outcomes. Many of the high-population outliers (shown in earlier sections) fail to integrate smoothly into their



(a) Clusters of the embedding from SR-BHGNN

(b) Clusters of the embedding from BHGNN

Figure 3. Comparison of spatial clustering in OA-level embedding clusters derived from (a) SR-BHGNN and (b) BHGNN. Maps share the same legend, north arrow, and scale bar. Embeddings are clustered using k-means ($k = 3$) and visualised with consistent colour assignments across panels.

surrounding areas, reflecting the absence of a mechanism to 'pull' nearby nodes closer in representation space. To further examine these differences, we applied t-SNE (Van der Maaten and Hinton 2008) to each model's embeddings, reducing their dimensionality to two dimensions. The t-SNE plots, as shown in Figure 3, reinforce that SR-BHGNN embeddings display markedly higher spatial autocorrelation, as quantified by a bivariate Moran's I of 0.53 ($p < 0.01$) (Lee 2001), indicating that physically adjacent OAs also cluster together in latent space.

The results presented in this subsection underscore the value of combining hierarchical graph structures, Bayesian parameter inference, and spatial regularisation in small-area population estimation. SR-BHGNN accommodates missing fine-scale features and manages a heavily imbalanced distribution more effectively than baseline approaches by systematically leveraging adjacency and membership relationships. In the next section, we will present another empirical analysis on air pollution prediction and demonstrate how uncertainty quantification can provide insights for a better understanding of urban environments.

4.3. Case study 2: predicting air pollution

Reliable, fine-scale estimates of air pollutant concentrations are crucial for urban planning, public health interventions, and policy-making, particularly in large and densely populated cities such as London (Walton *et al.* 2015, Maltby 2022). In this section, we extend the SR-BHGNN (with OA-level features) to predict PM $_{2.5}$ exposure (as introduced in Section 4.1) at the OA level, thereby illustrating the model's capacity to handle environmental data and facilitating uncertainty analysis for informed decision support. Although pollution levels are frequently recorded at specific monitoring stations, recent work shows that socio-economic and urban morphological variables can strongly influence pollutant generation, dispersion, and human exposure patterns (Rao *et al.* 2017, Yuan *et al.* 2024). Hence, we leverage LSOA-level socio-economic and morphological data to capture broader neighbourhood characteristics, while OA-level census attributes provide finer-grained demographic and housing information pertinent to local emission sources and vulnerabilities. By combining these multi-scale inputs, the proposed method can learn how macro-level conditions and micro-level features jointly shape $PM_{2.5}$ exposure, enabling accurate predictions and uncertainty estimates for air quality management.

Similar to the population classification experiment, we discretised the $PM_{2.5}$ exposure measurements into three categories (*low*, *median*, and *high*) using Jenks' natural breaks. The proposed SR-BHGNN model achieved strong classification performance, with an Accuracy of 0.81, a Precision of 0.78, a Recall of 0.80, and an F-score of 0.79. Beyond these classification metrics, we also assessed the quality of the model's predictive uncertainty using a multi-class generalisation of Continuous Ranked Probability Score (CRPS), a proper scoring rule widely used for probabilistic evaluation (Gneiting and Raftery 2007). As summarised in Table 3, SR-BHGNN outperformed all Bayesian baselines, achieving the lowest CRPS value (2.73), which indicates superior calibration of its predictive distributions. In contrast, BHGNN (without spatial regularisation) and Spatial HB yielded higher CRPS values of 4.91 and 3.87, respectively, suggesting less reliable uncertainty quantification. Notably, Spatial HB had better CRPS performance than BHGNN, despite showing lower accuracy and F-score. Such a discrepancy highlights a key distinction between accuracy (which measures the correctness of discrete predictions) and CRPS (which evaluates the calibration and reliability of probabilistic predictions). A model can thus produce less accurate point predictions but still deliver better-calibrated uncertainty estimates, particularly in ambiguous or spatially heterogeneous contexts, which are common in metropolitan areas like London. Hence, the results in Table 3 reinforce the importance of spatial regularisation within the hierarchical Bayesian GNN framework, as it simultaneously enhances classification accuracy and probabilistic calibration.

To further evaluate the spatial calibration of predictive uncertainty across London, we present both the distribution and spatial patterns of CRPS scores for the three models in Figure 5. The histogram (top left) shows that SR-BHGNN produces the lowest and most tightly distributed CRPS values, suggesting consistently well-calibrated predictions. In contrast, Spatial HB and BHGNN yield higher and more dispersed CRPS scores, reflecting weaker probabilistic calibration. The spatial maps further reveal that SR-BHGNN (top right) achieves lower uncertainty across most of the city, particularly in outer boroughs, while Spatial HB and BHGNN (bottom row) display widespread areas of elevated CRPS, indicating poor uncertainty reliability. Notably, even for SR-BHGNN, CRPS scores are relatively higher in several inner boroughs, which suggests that although the model performs well overall, capturing the probabilistic structure of

Table 3. Performance comparison of probabilistic models for PM $_{2.5}$ prediction.

Model	Accuracy	F-score	CRPS (↓)
SR-BHGNN	0.81	0.79	2.73
BHGNN	0.73	0.68	4.91
Spatial HB	0.65	0.61	3.87
RF	0.63	0.60	N/A
GRF	0.66	0.62	N/A
GAT	0.70	0.67	N/A

Lower CRPS values indicate better-calibrated predictive distributions.

complex central urban areas remains challenging, likely due to heightened socio-spatial heterogeneity. We examine such a spatial discrepancy in greater detail in the following paragraphs.

As shown in Figure 4 (left panels), the *ground-truth* PM $_{2.5}$ map (top left) is generally well aligned with the model's predicted labels (bottom-left). To better understand these results, we highlight on the right side the *mismatches* between the predicted and actual classes (top right, red areas) and the Bayesian model's *uncertainty* values (bottom-right, darker blue indicating higher uncertainty). Both mismatch and uncertainty exhibit significant global spatial clustering (Moran's $I = 0.38$ and 0.73, respectively, with $p < 0.01$), suggesting that errors and uncertain predictions are not randomly distributed but tend to cluster in specific regions (Figure 5).

Figure 6 provides additional insights into the spatial patterns of prediction mismatches and uncertainty, as well as their potential social implications. In panel (a), overlaid major roads (blue) reveal that the model's misclassification (red polygons) often cluster along key transport corridors. This finding suggests that relying solely on *Queen contiguity* to define adjacency in the graph may overlook crucial pathways of

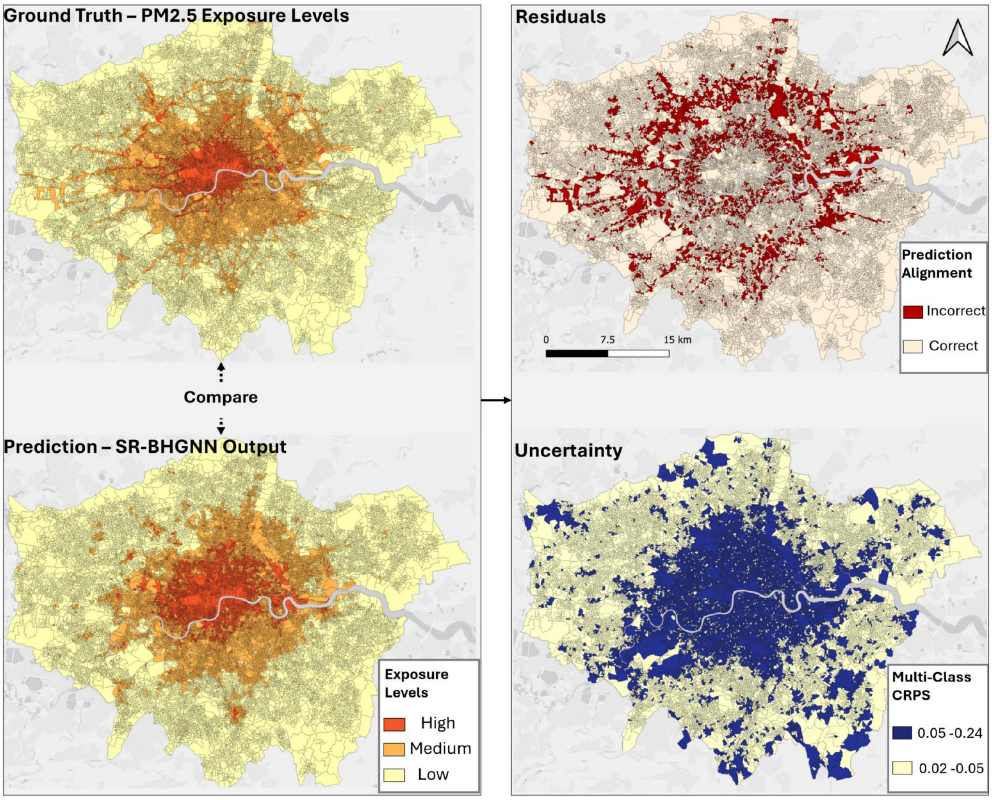


Figure 4. Observed and predicted OA-level PM $_{2.5}$ exposure classes, mismatches, and associated predictive uncertainty for the SR-BHGNN. All maps share the same north arrow and scale bar. Predicted exposure classes are derived from Jenks' natural breaks (three classes: low, medium, high) based on observed PM $_{2.5}$ values. Mismatch maps highlight OAs where predicted and observed classes differ (red), while uncertainty maps use darker shades to denote higher predictive variance.

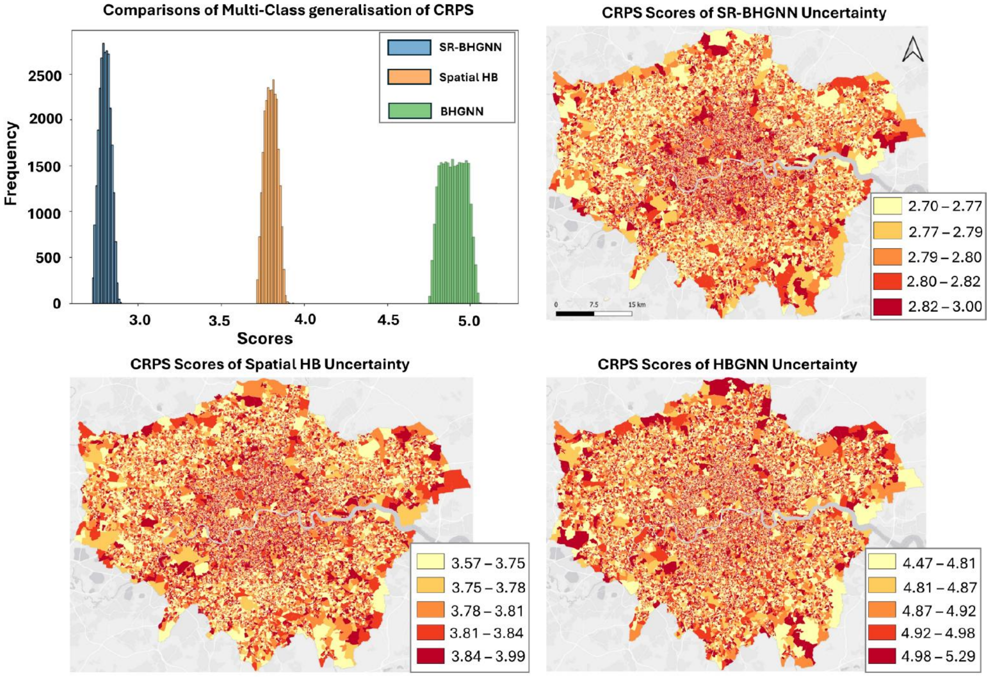


Figure 5. Comparison of Continuous Ranked Probability Score (CRPS) values across SR-BHGNN, Spatial HB, and BHGNN for PM $_{2.5}$ prediction. The top-left panel shows the distribution of CRPS scores; the top right and bottom panels show their spatial distribution. All maps share the same north arrow and scale bar. Note: each model's CRPS map uses classification breaks tailored to its CRPS range (SR-BHGNN: 2.70–3.00; Spatial HB: 3.57–3.99; BHGNN: 4.47–5.29). These classification differences should be considered when visually comparing models.

pollutant dispersion and concentration, which are known to be strongly influenced by traffic flows, street canyons, and road network connectivity (Shahid *et al.* 2021, Stucki *et al.* 2024). As such, we did another extra experiment to substitute a road-based adjacency definition for the default Queen contiguity in our heterogeneous GNN architecture resulted in a 5.32% increase in overall accuracy, alongside gains of 4.11% in precision, 3.98% in recall, and 5.01% in F-score, highlighting the tangible benefit of incorporating transportation networks more explicitly in PM $_{2.5}$ modelling.

Panel (b) highlights that high model uncertainty often coincides with the Inner London Lower Emission Zone (LEZ), pointing to the complex interplay of policy interventions and built-environment dynamics (Bosher *et al.* 2007, Fanzini and Venturini 2022). Despite the SR-BHGNN's overall ability to capture urban morphological features and socio-demographic characteristics, policy-induced variations (e.g. restrictions on certain vehicle types, newer bus fleets) may introduce rapidly shifting emission profiles that the model cannot fully account for, especially if it lacks fine-grained transport data. Rather than representing a failure, these zones of elevated uncertainty underscore the need for policy-aware AI modelling (De Falco and Romeo 2025), where knowledge of regulatory measures informs data collection (e.g. monitoring station placement) and model design. From a data complexity perspective, the target variable, PM $_{2.5}$ concentration, exhibits lower variability within the LEZ, as indicated by a lower

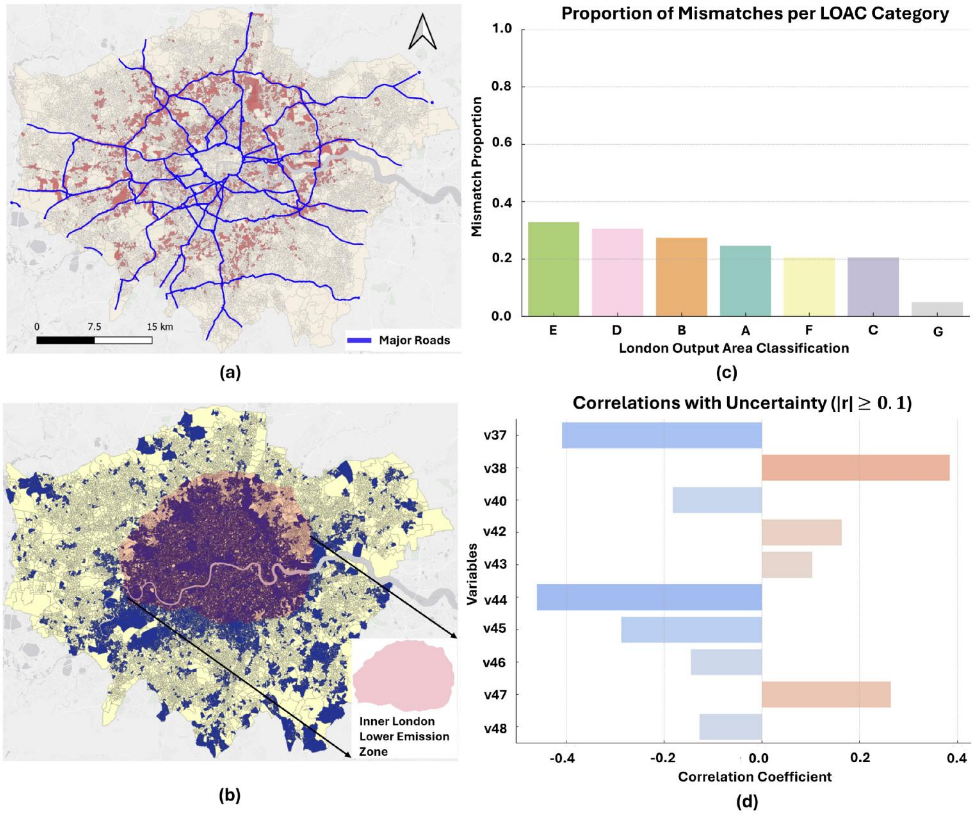


Figure 6. Further investigation on prediction mismatches and uncertainty. For Figure (c), A: Professional Employment and Family Lifecycles; B: The Greater London Mix; C: Suburban Asian Communities; D: Central Connected Professionals and Managers; E: Social Rented Sector Families with Children; F: Young Families and Mainstream Employment; G: Older Residents in Owner-Occupied Suburbs. For Figure (d), v37: Ownership or shared ownership; v38: Social rented; v40: Occupancy rating of rooms: +1 or more; v42: Standardised Disability Ratio; v43: Provides no unpaid care; v44: 2 or more cars or vans in household; v45: Highest level of qualification: Level 1, 2 or Apprenticeship; v46: Highest level of qualification: Level 3 qualifications; v47: Highest level of qualification: Level 4 qualifications or above; v48: Hours worked: Part-time.

local coefficient of variation ($\delta = 0.28$) compared to the rest of London ($\delta = 0.49$). Such a result suggests that areas with lower target complexity may pose greater challenges for the model because reduced variability can hinder the model's ability to learn robust predictive patterns, which ultimately affects the model's confidence when giving predictions.

Turning to panel (c), the bar chart reveals the proportion of mismatches by London Output Area Classification (LOAC) super group (Singleton and Longley 2024), indicating that areas typified by 'Social Rented Sector Families with Children' (Group E) experience the highest error rate. This pattern aligns with broader research on environmental justice, wherein low-income or socially rented communities often face higher pollution burdens and greater uncertainty in exposure estimates (Champion *et al.* 2022, Jbaily *et al.* 2022). Finally, panel (d) shows correlations between uncertainty and various socio-economic variables, including housing tenure and educational qualifications. Notably, a higher share of social renting (v38) correlates positively with uncertainty, while indicators of greater educational attainment or private car ownership relate negatively. Such correlations highlight the multi-dimensional drivers of model ambiguity: in some neighbourhoods, insufficient or fluctuating data on occupant behaviours and infrastructure may exacerbate uncertainty, whereas more stable or affluent communities tend to have better-captured emission patterns. These findings call for targeted data-collection efforts (e.g. improved road-based monitoring) and policy coordination that considers the heterogeneous nature of urban communities, ensuring that socially vulnerable populations are neither overlooked nor disproportionately impacted by modelling uncertainties.

The promising performance of SR-BHGNN in small-area population and PM $_{2.5}$ estimation tasks demonstrates the model's effectiveness in integrating hierarchical structures and spatial dependencies. For readers interested in a more detailed view of which predictors drive the model's outputs, Appendix C presents scale-specific feature-importance diagnostics for both case studies. These are provided as an interpretability aid only and are not discussed further in the main text. However, these analyses thus far have been confined to a two-tier census tract setup, that is, OAs nested within LSOAs. Urban systems, by contrast, often involve deeper hierarchies and broader regional influences. Hence, it raises an important question about whether the inclusion of additional census tract levels can further enhance prediction accuracy and uncertainty quantification. In what follows, we explore this question by extending our framework to incorporate a third census tract scale, Middle Layer Super Output Areas (MSOAs), and assess the incremental value it brings to fine-scale estimation tasks.

5. Unlocking the potential of hierarchy

Building on the motivation outlined at the end of Section 4.2, we now test a central premise in multi-level urban analytics: whether incorporating additional census tract tiers can provide valuable contextual information and thereby improve predictive performance and uncertainty estimation for lower-level geographies (Lloyd 2014). Extending the two-level design of the previous experiments (OAs nested within LSOAs), we integrate a third statistical geography scale – Middle Layer Super Output Areas (MSOAs) – as illustrated in Figure 7. Such a setup allows us to examine whether explicitly modelling deeper hierarchical structure enhances small-area population estimates beyond the gains achievable with two scales.

We draw upon the Climate Just (Climate Just 2014) project to incorporate 11 socio-environmental indicators at the MSOA level, covering a spectrum of vulnerability and adaptive capacity measures for flood and heat risks. These include, for instance, an area's 'sensitivity' to flooding, 'enhanced exposure' to heat, and a 'socio-spatial vulnerability index' for each hazard. In addition, we derive two supplementary metrics from OpenStreetMap (OSM) data, the aggregated number of health care facilities per MSOA and average addressable outdoor space (square metres) per MSOA, reflecting the accessibility of medical services and open spaces. Each MSOA is then assigned a vector of normalised features encompassing these 13 attributes. Similar to the experiment before, we define Queen contiguity edges among MSOAs to capture their adjacency

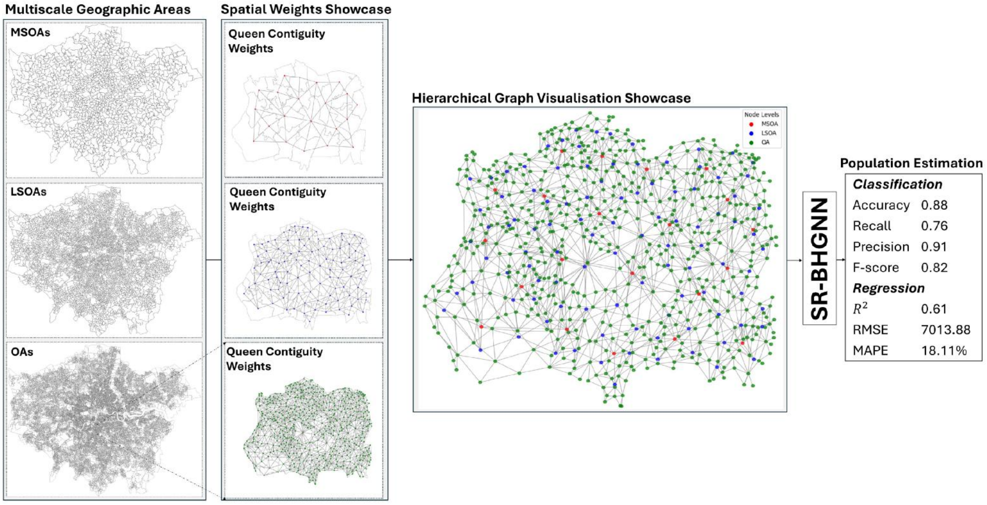


Figure 7. A combination of MSOA, LSOA and OA for the population estimation based on the SR-BHGNN framework.

relationships. The result is a three-tiered hierarchical structure: *MSOA* ⊃ *LSOAs* ⊃ *OAs*, where LSOAs remain nested within MSOAs, just as OAs remain nested within LSOAs. Using a methodology analogous to our earlier graphs, we link each LSOA node to exactly one MSOA node through 'membership' edges, thereby adding a new 'MSOA ↔ LSOA' layer to the heterogeneous graph. We then apply the SR-BHGNN to this augmented graph, training the model to predict population estimates at the OA level.

As shown in Figure 7, with the new MSOA layer, the *classification* task achieves an Accuracy of 0.88, a Recall of 0.76, a Precision of 0.91, and an F-score of 0.82. For the *regression* setting, the model attains an R^2 value of 0.61, an RMSE of 7013.88, and a MAPE of 18.11%. These results show notable improvement over the two-tier hierarchy (OAs nested within LSOAs) described in Section 4.2. Such gains in the model's performance may stem from the MSOA-level socio-environmental data, which supply broader contextual cues about infrastructural resources, climate vulnerabilities, and public health capacities, which are shaping residential patterns and community demographics in ways not fully captured by LSOA or OA attributes alone.

However, the magnitude of these improvements, while meaningful, is not transformative. Such a result suggests that simply 'stacking' additional census tract layers or data sources does not guarantee dramatic leaps in performance. The benefit appears to hinge on the relevance and distinctiveness of the new indicators. In some instances, MSOA-level indicators may echo patterns already visible at the LSOA level (e.g. similar demographic correlations or environmental risks), which may lead to diminishing returns (McMillen 2010).

Nevertheless, the findings in this section illustrate how introducing an additional mid-level scale can enhance small-area analytics and further validate the SR-BHGNN as a framework capable of 'unlocking the potential of scale'. By nesting MSOAs alongside existing LSOAs and OAs, the model gains deeper insights into broader regional contexts, such as climate vulnerability and healthcare infrastructure, while still capturing localised patterns. Although simply stacking more census tract layers does not guarantee major performance leaps, the moderate yet meaningful improvements in classification and regression emphasise that multi-scale integration remains a promising strategy for refining predictions and revealing cross-level interactions within complex urban environments (Weaver 2015).

6. Discussion

The SR-BHGNN proposed in this study aims to advance hierarchical spatial analytics and small-area estimation by seamlessly integrating multi-scale data into a unified, graph-based learning paradigm. Methodologically, the framework capitalises on recent developments in graph neural networks for non-Euclidean urban data representation (Liu and Biljecki 2022, Liu *et al.* 2025a), augmenting these with (i) Bayesian inference to capture parameter and predictive uncertainties, and (ii) heterogeneous message passing that respects both adjacency within each census tract layer and membership across scales (e.g. from OAs to LSOAs and vice versa). Such a design enables a principled means of 'borrowing strength' from higher-level aggregates when finer-scale observations are sparse, which is a core challenge in small-area problems (Banerjee *et al.* 2003, Rao and Molina 2015). By doing so, the SR-BHGNN differentiates itself from conventional single-level GNNs, which ignore hierarchical dependencies, and from standard hierarchical Bayesian models, which often impose strong parametric assumptions and lack robust ways of learning complex, non-linear spatial relationships.

A central innovation lies in the spatial regularisation term embedded in the loss function. Inspired by Tobler's First Law of Geography, which posits that spatial proximity correlates with higher similarity in observed phenomena (Tobler 1970, Anselin 2013), this term penalises large discrepancies in model predictions among adjacent areas unless strongly supported by local evidence. Empirical results across different experiments (e.g. population estimation, PM $_{2.5}$ pollution exposure) illustrate that omitting or weakening this regularisation leads to significant drops in model performance, particularly under data imbalance. For instance, in the population task, where a small number of OAs harbour disproportionately large resident densities, the lack of spatial smoothing compounds errors in outlier localities, degrading both classification accuracy and regression metrics. In contrast, enforcing spatial coherence helps capture the underlying spatial processes governing population or pollution distributions and reduces the adverse impact of extreme values.

Meanwhile, an important empirical finding emerges from the baseline comparisons. When higher-level features are artificially downscaled to OAs using area-weighted allocation and then concatenated with the original OA features, performance deteriorates across all metrics, whether in a neural model (GAT-Concat) or a non-parametric one (RF-Concat, GRF-Concat). Such results suggest that the challenge in multi-scale small-area estimation is not merely the availability of additional data, but rather the absence of a structural mechanism to regulate how information should flow across scales. In contrast, the hierarchical message passing in SR-BHGNN explicitly captures these cross-scale dependencies, allowing the model to exploit multi-scale signals without overwhelming the representation space or introducing pseudo–fine-scale noise (Goodchild and Lam 1980, Wong 2020). Consequently, the consistently poor performance of concatenation baselines serves as direct evidence against naïve feature fusion and provides strong validation for the architectural principles underpinning SR-BHGNN.

In terms of policy relevance, the SR-BHGNN model offers actionable insights for local decision-makers by producing granular, uncertainty-aware estimates that support equitable and data-informed resource allocation. In the context of population volume estimation, the model enables city councils, public health agencies, and urban planners to obtain reliable small-area estimates with credible intervals, which are crucial for planning services such as health clinics, social care facilities, and school capacities. Such benefits are particularly pronounced in areas with outdated census data or where rapid demographic shifts outpace official statistics (Longley *et al.* 2024). In the second application, predicting PM $_{2.5}$ exposure, the model effectively integrates morphological (e.g. road density) and socio-demographic (e.g. deprivation, housing tenure) variables to reflect both emission and dispersion dynamics. By capturing interdependencies across spatial tiers through a hierarchical graph, the model supports spatially adaptive policy design, for example, identifying hotspots that require the expansion of low-emission zones, prioritising neighbourhoods for new air quality monitors, or adjusting transport infrastructure in vulnerable districts. Crucially, the Bayesian design yields probabilistic predictions that are calibrated using CRPS, allowing policymakers to distinguish between areas with high predictive confidence and those where model uncertainty suggests the need for further investigation or data collection.

Moreover, further analysis indicates that introducing MSOAs, in addition to OAs and LSOAs, enriches the model with mid-scale socio-environmental indicators. This extra tier provides a more comprehensive view of how regional vulnerabilities and resources, including flood or heat sensitivity and healthcare infrastructure, can inform local population estimates. Although these enhancements do not radically transform predictive outcomes, they underline the synergy between finer and coarser geographies in capturing how macro-level factors (e.g. city-wide climate risks) intersect with more localised phenomena (e.g. neighbourhood demographics). By incorporating MSOA attributes into the heterogeneous graph, the SR-BHGNN captures more holistic cross-scale interactions that may be overlooked if only two census tract scales are considered. At the same time, these findings confirm that simply stacking additional layers does not guarantee substantial performance gains: the distinctiveness and relevance of the newly introduced features remain pivotal (Miller and Wentz 2003). Ultimately, blending mid-scale insights with fine-grained local data enables more nuanced analyses of urban processes, particularly in contexts where environmental risks or infrastructural disparities extend across multiple spatial tiers and shape local socio-demographics. In doing so, our proposed approach helps unlock the potential of multiscale urban analytics, resulting in a more holistic framework for interpreting and managing complex urban systems.

Although SR-BHGNN achieves strong performance across both case studies, the results should be interpreted in light of several data-related limitations. Our predictors draw on widely used, publicly available sources, Office for National Statistics small-area geographies, GSV and OSM, but each carries intrinsic constraints. OSM is a crowd-sourced product, and although coverage in London is generally high for the structural indicators used here, variations in completeness and positional accuracy by feature type or neighbourhood can introduce unquantified uncertainty.

A further limitation of this study concerns the temporal alignment of the datasets used. The Street View imagery spans 2019–2025, OSM features were extracted in 2025, and the PM $_{2.5}$ outcome for the second case study reflects conditions in 2013. The predictor set combines several types of variables with differing temporal sensitivities. Built-form and street-network indicators derived from OSM and GSV generally represent structural characteristics that tend to evolve gradually at the neighbourhood scale. In contrast, socio-demographic and deprivation measures (e.g. England IoD domains, LOAC variables) may respond more rapidly to policy interventions, modifications in service provision, or changes in demographic composition in the UK context (Singleton and Longley 2009, Singleton *et al.* 2016). Because the degree of local change during these intervals cannot be fully established with the available data, the temporal mismatch between predictors and outcomes may contribute to the uncertainties observed in some areas.

Importantly, this study is primarily methodological in nature, and we did not attempt comprehensive temporal harmonisation or a formal audit of neighbourhood-level change. We therefore treat temporal misalignment as an inherent source of uncertainty that may affect different predictors to different extents, without assuming uniform temporal stability across the feature set. Future applications of SR-BHGNN, particularly in rapidly changing urban contexts or when using more time-sensitive predictors, can benefit from a deeper assessment of temporal sensitivity, data vintage, and their implications for multi-scale modelling.

7. Conclusion

This study introduces SR-BHGNN, a novel spatially regularised Bayesian hierarchical graph neural network for small-area estimation. By combining non-linear message passing, hierarchical pooling, and spatial regularisation, the model captures the complex interdependencies that shape urban phenomena across multiple census tract levels. Empirical evaluations on population estimation and air pollution prediction demonstrate that SR-BHGNN improves predictive accuracy and calibration, particularly in imbalanced or noisy datasets, while also providing actionable uncertainty estimates.

Future research could explore dynamic or temporal extensions to incorporate shifting demographic patterns or seasonal variations in pollutant concentrations, thereby capturing the evolving nature of urban systems. Alternative adjacency definitions, such as road-centric or multi-layer topologies, might yield superior results in contexts where linear infrastructure determines movement patterns and emissions. Scenario analyses that manipulate policy assumptions or the degree of data availability would further clarify the robustness of model outcomes. Incorporating richer policy metadata, such as congestion charges, emission standards, or green infrastructure investments, may also enhance predictive accuracy and relevance by reflecting real-world regulatory levers.

Acknowledgments

The authors thank the editor and the reviewers for their constructive, thoughtful, and detailed feedback, which significantly improved the quality of this manuscript. The research was partly conducted at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and the National Research Foundation Singapore.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This research is supported by the National Research Foundation Singapore (NRF) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The second author, Yang Chen, is supported by the China Scholarship Council (No. 202406860051). The third author, Xiucheng Liang, acknowledges the NUS Graduate Research Scholarship granted by the National University of Singapore (NUS). We express our gratitude to the members of the NUS Urban Analytics Lab for their valuable discussions and insights.

Notes on contributors

Pengyuan Liu is a Lecturer in Digital Planning at the University of Glasgow. He holds an MSc in Cloud Computing and a PhD in Geography from the University of Leicester, United Kingdom.

Yang Chen is a PhD student from Nanjing Normal University and a visiting PhD scholar at the NUS Urban Analytics Lab.

Xiucheng Liang is a PhD student at the NUS Urban Analytics Lab. He holds an MSc in Architecture from the National University of Singapore.

Hao Li is a Lecturer at the Department of Geography, National University of Singapore. He holds an MSc in Geomatics and a PhD in GIS from Heidelberg University, Germany. His research interests include GIScience, GeoAI, spatial computing and VGI.

Filip Biljecki is an Assistant Professor at the National University of Singapore and the principal investigator of the NUS Urban Analytics Lab. He holds an MSc in Geomatics and a PhD in 3D GIS from the Delft University of Technology in the Netherlands.

Rudi Stouffs is Dean's Chair Associate Professor in the Department of Architecture and Assistant Dean (Research) in the College of Design and Engineering, National University of Singapore. He received his PhD in Architecture from Carnegie Mellon University and has held previous appointments at Carnegie Mellon University, ETH Zurich, and TU Delft.

ORCID

Pengyuan Liu http://orcid.org/0000-0002-5443-5910
Yang Chen http://orcid.org/0000-0003-1283-4363
Xiucheng Liang http://orcid.org/0000-0003-0898-7543
Hao Li http://orcid.org/0000-0002-6336-8772
Filip Biljecki http://orcid.org/0000-0002-6229-7749
Rudi Stouffs http://orcid.org/0000-0002-4200-5833

Data and codes availability statement

Due to commercial licencing restrictions of the Google Street View imagery, the full dataset cannot be shared. Instead, we provide a representative subset of the data, together with the end-to-end analytical pipelines and code for recreating the tables and figures in the paper at: https://doi.org/10.6084/m9.figshare.29856452. This subset is sufficient to reproduce the complete analysis workflow, including all pre-processing, modelling, and visualisation steps, although the specific numerical values in the tables and figures will differ from those reported in the manuscript.

References

Ansel, J., *et al.*, 2024. PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation. *In*: 29th ACM international conference on architectural support for programming languages and operating systems, volume 2 (ASPLOS '24). New York: ACM. Available from: https://pytorch.org/assets/pytorch2-2.pdf.

Anselin, L., 1988. *Spatial econometrics: methods and models*. Vol. 4. Dordrecht, Netherlands: Springer Science & Business Media.

Anselin, L., 2013. *Spatial econometrics: methods and models*. Vol. 4. Dordrecht, Netherlands: Springer Science & Business Media.

Arribas-Bel, D. and Fleischmann, M., 2022. Spatial signatures-understanding (urban) spaces through form and function. *Habitat International*, 128, 102641.

Atkinson, P.M., Pardo-Iguzquiza, E., and Chica-Olmo, M., 2008. Downscaling cokriging for super-resolution mapping of continua in remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 46 (2), 573–580.

Atkinson, P.M. and Tate, N.J., 2000. Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, 52 (4), 607–623.

Bajjali, W., 2023. Spatial interpolation. *In*: *Arcgis pro and arcgis online: applications in water and environmental sciences*. Cham, Switzerland: Springer, 223–242.

Banerjee, S., Carlin, B.P., and Gelfand, A.E., 2003. *Hierarchical modeling and analysis for spatial data*. New York: Chapman and Hall/CRC.

Barrington-Leigh, C. and Millard-Ball, A., 2020. Global trends toward urban street-network sprawl. *Proceedings of the National Academy of Sciences*, 117 (4), 1941–1950.

Berild, M.O., *et al.*, 2022. Importance sampling with the integrated nested laplace approximation. *Journal of Computational and Graphical Statistics*, 31 (4), 1225–1237.

Besag, J., York, J., and Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43 (1), 1–20.

Biljecki, F., Chow, Y.S., and Lee, K., 2023. Quality of crowdsourced geospatial building information: a global assessment of openstreetmap attributes. *Building and Environment*, 237, 110295.

Biljecki, F. and Ito, K., 2021. Street view imagery in urban analytics and gis: a review. *Landscape and Urban Planning*, 215, 104217.

Bingham, E., *et al.*, 2019. Pyro: deep universal probabilistic programming. *Journal of Machine Learning Research*, 20, 28:1–28:6. http://jmlr.org/papers/v20/18-403.html.

Blei, D.M., Kucukelbir, A., and McAuliffe, J.D., 2017. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112 (518), 859–877.

Blundell, C., *et al.*, 2015. Weight uncertainty in neural network. In: *International conference on machine learning*. Lille, France: PMLR, 1613–1622.

Boeing, G., 2017. Osmnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126–139.

Boeing, G., 2022. Street network models and indicators for every urban area in the world. *Geographical Analysis*, 54 (3), 519–535.

Bosher, L., *et al.*, 2007. Realising a resilient and sustainable built environment: towards a strategic agenda for the united kingdom. *Disasters*, 31 (3), 236–255.

Briggs, D.J., *et al.*, 2007. Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sensing of Environment*, 108 (4), 451–466.

Britten, G.L., *et al.*, 2021. Evaluating the benefits of bayesian hierarchical methods for analyzing heterogeneous environmental datasets: a case study of marine organic carbon fluxes. *Frontiers in Environmental Science*, 9, 491636.

Cao, C. and Su, Y., 2024. Transportation infrastructure and regional resource allocation. *Cities*, 155, 105433.

Champion, W.M., Khaliq, M., and Mihelcic, J.R., 2022. Advancing knowledge to reduce lead exposure of children in data-poor low-and middle-income countries. *Environmental Science & Technology Letters*, 9 (11), 879–888.

Chen, D., *et al.*, 2024. Bayesian hierarchical graph neural networks with uncertainty feedback for trustworthy fault diagnosis of industrial processes. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (12), 18635–18648.

Cheng, B., *et al.*, 2022. Masked-attention mask transformer for universal image segmentation. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1290–1299. Silver Spring, MD: IEEE Computer Society.

Cheng, F., *et al.*, 2024. Spatial downscaling of downward surface shortwave radiation based on image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–11.

Chi, G. and Zhu, J., 2008. Spatial regression models for demographic analysis. *Population Research and Policy Review*, 27 (1), 17–42.

Chiles, J.P. and Delfiner, P., 2012. *Geostatistics: modeling spatial uncertainty*. Hoboken, NJ: John Wiley & Sons.

Clayton, D. and Kaldor, J., 1987. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43 (3), 671–681.

Clifton, K., *et al.*, 2008. Quantitative analysis of urban form: a multidisciplinary review. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 1 (1), 17–45.

Climate Just, 2014. Climate just: an information tool designed to help with the delivery of equitable responses to climate change at the local level. Available from: https://www.climatejust.org.uk/ [Accessed 7 Apr 2025].

Corral, P., *et al.*, 2022. *Guidelines to small area estimation for poverty mapping*. Washington: World Bank.

Cressie, N., 2015. *Statistics for spatial data*. Hoboken, NJ: John Wiley & Sons.

De Falco, C.C. and Romeo, E., 2025. Algorithms and geo-discrimination risk: what hazards for smart cities' development? *In*: *Smart cities*. Abingdon, Oxfordshire, UK: Routledge, 104–117.

De Sabbata, S. and Liu, P., 2023. A graph neural network framework for spatial geodemographic classification. *International Journal of Geographical Information Science*, 37 (12), 2464–2486.

Demšar, U., *et al.*, 2013. Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, 103 (1), 106–128.

Diaconescu, A.O., *et al.*, 2014. Inferring on the intentions of others by hierarchical bayesian learning. *PLoS Computational Biology*, 10 (9), e1003810.

Dong, G. and Harris, R., 2015. Spatial autoregressive models for geographically hierarchical data structures. *Geographical Analysis*, 47 (2), 173–191.

Edochie, I., *et al.*, 2025. Small area estimation of poverty in four west african countries by integrating survey and geospatial data. *Journal of Official Statistics*, 41 (1), 96–124.

Fan, Z., *et al.*, 2023. Urban visual intelligence: uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences of the United States of America*, 120 (27), e2220417120.

Fanzini, D. and Venturini, G., 2022. Intervention on the built environment: approaches and strategies for the city. In: *Reactivation of the built environment: from theory to practice*. Cham: Springer, 1–14.

Ferrario, E., Pedroni, N., and Zio, E., 2016. Evaluation of the robustness of critical infrastructures by hierarchical graph representation, clustering and monte carlo simulation. *Reliability Engineering & System Safety*, 155, 78–96.

Fey, M. and Lenssen, J.E., 2019. Fast graph representation learning with PyTorch Geometric. *In*: *ICLR workshop on representation learning on graphs and manifolds, New Orleans, USA.*

Fotheringham, A.S., 2024. How to solve the scale "problem" in spatial analytics. In: *A research agenda for spatial analysis*. Cheltenham, UK: Edward Elgar Publishing, 55–66.

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2009. Geographically weighted regression. *The Sage Handbook of Spatial Analysis*, 1, 243–254.

Fuglstad, G.A., *et al.*, 2014. Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25, 115–133.

Gao, K., *et al.*, 2023. Data-driven interpretation on interactive and nonlinear effects of the correlated built environment on shared mobility. *Journal of Transport Geography*, 110, 103604.

Georganos, S. and Kalogirou, S., 2022. A forest of forests: a spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS International Journal of Geo-Information*, 11 (9), 471.

Geyer, C.J., 1992. Practical markov chain monte carlo. *Statistical Science*, 7 (4), 473–483.

Ghosh, M., 2021. *Bayesian methods for finite population sampling*. Abingdon, Oxon, UK: Routledge.

Ghosh, M. and Rao, J.N., 1994. Small area estimation: an appraisal. *Statistical Science*, 9 (1), 55–76.

Ghosh, S., *et al.*, 2024. Graph theory applications for advanced geospatial modelling and decision-making. *Applied Geomatics*, 16 (4), 799–812.

Gneiting, T. and Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102 (477), 359–378.

Gocht, A. and Röder, N., 2014. Using a bayesian estimator to combine information from a cluster analysis and remote sensing data to estimate high-resolution data for agricultural production in germany. *International Journal of Geographical Information Science*, 28 (9), 1744–1764.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.

Goodchild, M.F. and Lam, N.S.N., 1980. Areal interpolation: a variant of the traditional spatial problem. *Geo-processing*, 1 (3), 297–312.

Google LLC, 2025. Google maps/google street view apis terms of service. Available from: https://developers.google.com/maps/terms [Accessed 13 Feb 2025].

Gottdiener, M.D., Lehtovuori, P., and Budd, L., 2015. Key concepts in urban studies. 2nd ed. *SAGE key concepts series*. London, UK: SAGE Publications. Available from: https://uk.sagepub.com/en-gb/eur/key-concepts-in-urban-studies/book234192

Greater London Authority, 2017a. Average house prices by Borough, Ward, MSOA & LSOA. Available from: https://data.london.gov.uk/dataset/average-house-prices [Accessed 2 Apr 2025].

Greater London Authority, 2017b. Pm2.5 map and exposure data. Available from: https://data.london.gov.uk/dataset/pm2-5-map-and-exposure-data [Accessed 3 Apr 2025].

Guan, Q., Kyriakidis, P.C., and Goodchild, M.F., 2011. A parallel computing approach to fast geostatistical areal interpolation. *International Journal of Geographical Information Science*, 25 (8), 1241–1267.

Huang, H. and Abdel-Aty, M., 2010. Multilevel data and bayesian analysis in traffic safety. *Accident; Analysis and Prevention*, 42 (6), 1556–1565.

Ito, K., *et al.*, 2025. Zensvi: an open-source software for the integrated acquisition, processing and analysis of street view imagery towards scalable urban science. *Computers, Environment and Urban Systems*, 119, 102283.

Jbaily, A., *et al.*, 2022. Air pollution exposure disparities across us population and income groups. *Nature*, 601 (7892), 228–233.

Jenks, G.F., 1967. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7, 186–190.

Jiang, J. and Rao, J.S., 2020. Robust small area estimation: an overview. *Annual Review of Statistics and Its Application*, 7 (1), 337–360.

Kingma, D.P., Salimans, T., and Welling, M., 2015. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, 28, 2575–2583.

Kirilenko, A.P., 2022. Geographic information system (gis) making sense of geospatial data. In: *Applied data science in tourism: interdisciplinary approaches, methodologies, and applications*. Cham: Springer, 513–526.

Kullback, S. and Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1), 79–86.

Lee, S.I., 2001. Developing a bivariate spatial association measure: an integration of pearson's r and moran's i. *Journal of Geographical Systems*, 3 (4), 369–385.

LeSage, J. and Pace, R.K., 2009. *Introduction to spatial econometrics*. Boca Raton, FL: Chapman and Hall/CRC.

Li, H., *et al.*, 2023a. Semi-supervised learning from Street-View images and OpenStreetMap for automatic building height estimation. *In*: 12th international conference on geographic information science (GIScience 2023). Wadern, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 7:1–7:15.

Li, S., *et al.*, 2023b. A hierarchical constraint-based graph neural network for imputing urban area data. *International Journal of Geographical Information Science*, 37 (9), 1998–2019.

Liu, F., *et al.*, 2014. Bayesian regularization via graph laplacian. *Bayesian Analysis*, 9 (2), 449–474.

Liu, P., 2024. *Spatial analysis*. Cham: Springer International Publishing.

Liu, P. and Biljecki, F., 2022. A review of spatially-explicit geoai applications in urban geography. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102936.

Liu, P. and De Sabbata, S., 2021. A graph-based semi-supervised approach to classification learning in digital geographies. *Computers, Environment and Urban Systems*, 86, 101583.

Liu, P., *et al.*, 2025a. Sensing climate justice: a multi-hyper graph approach for classifying urban heat and flood vulnerability through street view imagery. *Sustainable Cities and Society*, 118, 106016.

Liu, P., *et al.*, 2025b. Living upon networks: a heterogeneous graph neural embedding integrating waterway and street systems for urban form understanding. *Environment and Planning B: Urban Analytics and City Science*.

Lloyd, C.D., 2014. *Exploring spatial scale in geography*. Chichester, West Sussex, UK: John Wiley & Sons.

Longley, P., Lan, T., and van Dijk, J., 2024. Geography, ethnicity, genealogy and intergenerational social inequality in great britain. *Transactions of the Institute of British Geographers*, 49 (1), e12622.

Luo, P., *et al.*, 2025. Measuring univariate effects in the interaction of geographical patterns. *International Journal of Geographical Information Science*, 1–32.

Mai, G., *et al.*, 2024. Srl: towards a general-purpose framework for spatial representation learning. *In: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*. 465–468.

Maltby, T., 2022. Consensus and entrepreneurship: the contrasting local and national politics of uk air pollution. *Environment and Planning C: Politics and Space*, 40 (3), 685–704.

Maxwell, K., Rajabi, M., and Esterle, J., 2021. Spatial interpolation of coal properties using geographic quantile regression forest. *International Journal of Coal Geology*, 248, 103869.

McDonald, J.F., 1989. Econometric studies of urban population density: a survey. *Journal of Urban Economics*, 26 (3), 361–385.

McMillen, D.P., 2010. Issues in spatial data analysis. *Journal of Regional Science*, 50 (1), 119–141.

Miller, H.J. and Wentz, E.A., 2003. Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, 93 (3), 574–594.

Molina, I., Nandram, B., and Rao, J.N.K., 2014. Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *The Annals of Applied Statistics*, 8 (2), 852–885.

Morales, D., *et al.*, 2021. *A course on small area estimation and mixed models.* Methods, *theory and applications in R*. Cham, Switzerland: Springer.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1-2), 17–23.

Mu, L. and Wang, X., 2006. Population landscape: a geometric approach to studying spatial patterns of the us urban hierarchy. *International Journal of Geographical Information Science*, 20 (6), 649–667.

Neuhold, G., *et al.*, 2017. The mapillary vistas dataset for semantic understanding of street scenes. In: *Proceedings of the IEEE international conference on computer vision*. 4990–4999. Piscataway, NJ: IEEE Computer Society.

Newbold, K.B., 2021. *Population geography: tools and issues*. Lanham, MD: Rowman & Littlefield.

Office for National Statistics, 2021. Census data. Available from: https://www.ons.gov.uk/census [Accessed 2 Apr 2025].

Oliver, M.A. and Webster, R., 1990. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4 (3), 313–332.

Pedregosa, F., *et al.*, 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Pfeffermann, D., 2002. Small area estimation-new developments and directions. *International Statistical Review/Revue Internationale de Statistique*, 70 (1), 125–143.

Rampichini, C. and Schifini d'Andrea, S., 1998. A hierarchical ordinal probit model for the analysis of life satisfaction in italy. *Social Indicators Research*, 44 (1), 41–69.

Rao, J.N. and Molina, I., 2015. *Small area estimation*. Hoboken, NJ: John Wiley & Sons.

Rao, S., *et al.*, 2017. Future air pollution in the shared socio-economic pathways. *Global Environmental Change*, 42, 346–358.

Rey, S.J., *et al.*, 2022. The pysal ecosystem: philosophy and implementation. *Geographical Analysis*, 54 (3), 467–487.

Risser, M.D. and Calder, C.A., 2015. Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26 (4), 284–297.

Romero, D., Ma, M., and Giannakis, G.B., 2017. Kernel-based reconstruction of graph signals. *IEEE Transactions on Signal Processing*, 65 (3), 764–778.

Rue, H., Martino, S., and Chopin, N., 2009. Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71 (2), 319–392.

Schnake-Mahl, A.S., *et al.*, 2020. Gentrification, neighborhood change, and population health: a systematic review. *Journal of Urban Health: bulletin of the New York Academy of Medicine*, 97 (1), 1–25.

Sekulić, A., *et al.*, 2020. Random forest spatial interpolation. *Remote Sensing*, 12 (10), 1687.

Shahid, N., *et al.*, 2021. Towards greener smart cities and road traffic forecasting using air pollution data. *Sustainable Cities and Society*, 72, 103062.

Shi, L., Yang, S., and Gao, L., 2016. Effects of a compact city on urban resources and environment. *Journal of Urban Planning and Development*, 142 (4), 05016002.

Singleton, A., Alexiou, A., and Savani, R., 2020. Mapping the geodemographics of digital inequality in great britain: an integration of machine learning into small area estimation. *Computers, Environment and Urban Systems*, 82, 101486.

Singleton, A.D. and Longley, P.A., 2024. Classifying and mapping residential structure through the london output area classification. *Environment and Planning B: Urban Analytics and City Science*, 51 (5), 1153–1164.

Singleton, A., Pavlis, M., and Longley, P.A., 2016. The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems*, 18 (2), 97–123.

Singleton, A.D. and Longley, P.A., 2009. Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29 (3), 289–298.

Sinha, P., *et al.*, 2019. Assessing the spatial sensitivity of a random forest model: application in gridded population modeling. *Computers, Environment and Urban Systems*, 75, 132–145.

Stucki, L., *et al.*, 2024. Long-term exposure to air pollution, road traffic noise and greenness, and incidence of myocardial infarction in women. *Environment International*, 190, 108878.

Sun, Y., 1971. Population estimation. *Annual Convention of the CC*, 4, 615.

Tate, N. and Atkinson, P.M., 2001. *Modelling scale in geographical information science*. Chichester, UK: John Wiley & Sons.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46 (sup1), 234–240.

Tong, Z., *et al.*, 2022. Exploring non-linear and spatially non-stationary relationships between commuting burden and built environment correlates. *Journal of Transport Geography*, 104, 103413.

Trust for London, 2019. Index of multiple deprivation 2019 (rebased) – London. Available from: https://trustforlondon.org.uk/data/index-multiple-deprivation-2019-rebased-london/ [Accessed 2 Apr 2025].

Tümtürk, O., Karakiewicz, J., and de Haan, F.J., 2024. Measuring the impact of plot types on physical change: a diachronic analysis of urban form evolution in new york, melbourne and barcelona. *Cities*, 154, 105380.

Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (11), 2579–2605.

Vatcheva, K.P., *et al.*, 2016. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, 6 (2), 227.

Velickovic, P., *et al.*, 2017. Graph attention networks. *stat*, 1050 (20), 10–48550.

Wall, M.M., 2004. A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121 (2), 311–324.

Walton, H., *et al.*, 2015. Understanding the health impacts of air pollution in london. *London: Kings College London, Transport for London and the Greater London Authority*, 1 (1), 6–14.

Wang, H., *et al.*, 2021. Hierarchical visualization of geographical areal data with spatial attribute association. *Visual Informatics*, 5 (3), 82–91.

Wang, P., Bayram, B., and Sertel, E., 2022. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232, 104110.

Wang, S., *et al.*, 2024. Mapping the landscape and roadmap of geospatial artificial intelligence (geoai) in quantitative human geography: an extensive systematic review. *International Journal of Applied Earth Observation and Geoinformation*, 128, 103734.

Wang, Y. and Zhu, D., 2024. A hypergraph-based hybrid graph convolutional network for intracity human activity intensity prediction and geographic relationship interpretation. *Information Fusion*, 104, 102149.

Weaver, R., 2015. A cross-level exploratory analysis of "neighborhood effects" on urban behavior: an evolutionary perspective. *Social Sciences*, 4 (4), 1046–1066.

Whitworth, A., *et al.*, 2017. Estimating uncertainty in spatial microsimulation approaches to small area estimation: a new approach to solving an old problem. *Computers, Environment and Urban Systems*, 63, 50–57.

Wikle, C.K., Berliner, L.M., and Cressie, N., 1998. Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, 5 (2), 117–154.

Wolf, L.J., *et al.*, 2021. On spatial and platial dependence: examining shrinkage in spatially dependent multilevel models. *Annals of the American Association of Geographers*, 111 (6), 1–13.

Wong, D., 2020. Aggregation effects in geo-referenced data. In: *Practical handbook of spatial statistics*. Boca Raton, FL: CRC Press, 83–106.

Wyszomierski, J., *et al.*, 2024. A neighbourhood output area classification from the 2021 and 2022 uk censuses. *The Geographical Journal*, 190 (2), e12550.

Yao, Y., *et al.*, 2017. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science*, 31 (6), 1–25.

Yap, W. and Biljecki, F., 2023. A global feature-rich network dataset of cities and dashboard for comprehensive urban analyses. *Scientific Data*, 10 (1), 667.

Yuan, Z., *et al.*, 2024. Hyperlocal air pollution mapping: a scalable transfer learning lur approach for mobile monitoring. *Environmental Science & Technology*, 58 (32), 14372–14383.

Yue, L., *et al.*, 2015. Fusion of multi-scale dems using a regularized super-resolution method. *International Journal of Geographical Information Science*, 29 (12), 2095–2120.

Zhang, D., 2019. Bayesian classification. In: *Fundamentals of image data mining: Analysis, features, classification and retrieval*. Cham, Switzerland: Springer, 161–178.

Zhu, D. and Ma, Z., 2025. Gravity-informed deep flow inference for spatial evolution modeling in panel data. *International Journal of Geographical Information Science*, 1–29.

Zhu, D., *et al.*, 2020. Understanding place characteristics in geographic contexts through graph convolutional neural networks. *Annals of the American Association of Geographers*, 110 (2), 408–420.

Zoraghein, H. and Leyk, S., 2018. Enhancing areal interpolation frameworks through dasymetric refinement to create consistent population estimates across censuses. *International Journal of Geographical Information Science: IJGIS*, 32 (10), 1948–1976.

Appendix A. Sensitivity analyses: graph construction and additional baselines

This appendix reports additional analyses of SR-BHGNN under alternative graph specifications and with further baseline models. For consistency across methods, we use Case Study 1 (population estimation) as the basis for these comparisons. In addition to the Queen continuity spatial weights used in the main text for OA- and LSOA-level graph construction, in this appendix, we further explored the following commonly used adjacency definitions for census tracts:

- Rook contiguity
- Centroid-based k-nearest neighbours (kNN): $k = \{8, 10, 12, 14, 16\}$
- Centroid distance bands: $d \in \{500, 1000, 2000\}$ metres

The cross-level connection between OAs and LSOAs remains the same as described in Section 3 in the main text. Because our study operates on census units rather than street-level topologies, barrier-aware or street-network-based graphs (e.g. removing edges across rivers or modelling bridges) are not implemented here and remain outside the scope of this paper.

In addition to RF, GAT, GAT-Concat, and Spatial HB reported in the main text, we include:

- *Besag–York–Mollié (BYM2) model with Integrated Nested Laplace Approximation (INLA):* The BYM2 model is a classic spatial regression method used in disease mapping and small-area estimation (Besag *et al.* 1991). It decomposes spatial random effects into two parts: a structured component that follows an intrinsic CAR, capturing spatial dependence, and an unstructured, independent and identically distributed element to account for overdispersion. We adapt the model to accommodate different likelihoods (a Poisson likelihood for regression and a multinomial–logit likelihood for classification) and fit it using INLA (Rue *et al.* 2009), with default penalised-complexity priors on the precision parameters. This provides a widely used Bayesian benchmark distinct from our Gibbs-based Spatial HB implementation.
- *Graph Laplacian–regularised ridge (GLR):* a non-deep, transductive linear baseline minimising $\| y - Xw \|_2^2 + \alpha \| w \|_2^2 + \beta\, \mathbf{f}^\top L\, \mathbf{f}$, where $\mathbf{f} = Xw$ are node-level predictions and L is the graph Laplacian (on Queen contiguity); this enforces smoothness of predictions over the graph without learned message passing (Liu *et al.* 2014, Romero *et al.* 2017). We tune the regularisation parameters α and β on a small validation grid and report the best-performing values.

The training and test splits used for the BYM/INLA and GLR baselines are identical to those used for SR-BHGNN and baselines, ensuring that all models are evaluated on the same data.

Appendix Table A1. Sensitivity of SR-BHGNN to graph construction and comparison to additional baselines.

Graph Variant	Acc	Prec	Rec	F1	R^2	RMSE	MAPE (%)
Rook	0.81	0.77	0.76	0.77	0.55	7611.45	41.25
kNN, $k = 8$	0.80	0.76	0.75	0.76	0.54	7812.67	44.04
kNN, $k = 10$	0.81	0.77	0.76	0.77	0.56	7553.46	40.59
kNN, $k = 12$	0.81	0.77	0.76	0.77	0.55	7792.12	41.32
kNN, $k = 14$	0.80	0.76	0.75	0.76	0.54	7901.23	41.81
kNN, $k = 16$	0.70	0.67	0.68	0.68	0.50	9467.90	49.50
Distance band, $d = 500$ m	0.79	0.75	0.74	0.74	0.52	8112.66	42.83
Distance band, $d = 1000$ m	0.77	0.75	0.74	0.75	0.51	8209.28	43.05
Distance band, $d = 2000$ m	0.69	0.66	0.71	0.69	0.50	9678.47	48.57
Baseline Model							
BYM2 (INLA)	0.60	0.58	0.56	0.57	0.43	12451.97	57.81
GLR	0.64	0.60	0.57	0.58	0.36	13751.63	63.57
RF-Concat	0.49	0.48	0.45	0.46	0.04	21186.21	137.91
GRF-Concat	0.39	0.33	0.32	0.33	0.01	21891.43	154.72

Classification metrics (left) correspond to the population classification task; regression metrics (right) correspond to continuous population regression outcomes.

Specifically, the BYM2 model utilises the same input as RF, GRF, and GAT; GLR uses the same inputs as for SR-BHGNN and BHGNN. Appendix Table A1 summarises indicative performance for the classification task (Accuracy/Precision/Recall/F1) and regression task (R^2/RMSE/MAPE). Moreover, as described in Section 4.2, we report the results of RF-Concat and GRF-Concat in this appendix as additional baselines.

Across all alternative adjacency specifications in Appendix Table A1, SR-BHGNN maintains performance close to that reported in Table 2 in the main text. Results under Rook contiguity and moderate k-nearest-neighbour or distance-band graphs differ only marginally, indicating that the model is not unduly sensitive to reasonable choices of within-scale graph construction. Echoing existing literature on urban graph modelling with GNNs (Liu and De Sabbata 2021), extensive neighbourhood definitions (e.g. $k = 16$ or a 2 km distance band) slightly reduce predictive power by over-smoothing.

The additional baselines, BYM2 (INLA) and GLR, perform in line with the main-text comparators: both improve on a purely aspatial random forest but remain well below SR-BHGNN on both classification and regression metrics, which reinforces the main-text finding that explicitly modelling hierarchical structure and spatial regularisation within a learned representation delivers consistent gains. Meanwhile, the results of RF-Concat and GRF-Concat support our finding that simply downscaling higher-level features and appending them to OA covariates does not improve accuracy; instead, it introduces pseudo–fine-scale noise (Goodchild and Lam 1980, Wong 2020) and undermines model stability.

Appendix B. Spatially robust validation

To complement the random 70/30 splits used throughout the main text, we performed a spatially robust validation to assess the impact of train–test proximity. Specifically, we implemented a *spatial block cross-validation* scheme at the OA level, still using Case Study 1 here to demonstrate the model performance.

London was partitioned into K non-overlapping, contiguous spatial blocks (based on k-means clustering on OA centroids in projected space, $K = 5$). In each fold, one block was held out for testing while the remaining blocks were used for training, ensuring that no OA in the test fold is spatially adjacent to any OA in the training data. This reduces leakage due to train–test proximity and provides a stricter estimate of out-of-sample performance.

Hyperparameters were kept identical to those in the main experiments. For each fold, we recorded classification (Accuracy, Precision, Recall, F1) and regression (R^2, RMSE, MAPE) metrics and then averaged across folds.

The spatially robust validation yields slightly lower metrics, as expected under stricter out-of-sample conditions; however, it demonstrates that our model still performs well under a spatially robust split, showing that the gains observed in the main text are not an artefact of train–test proximity (Appendix Table B1).

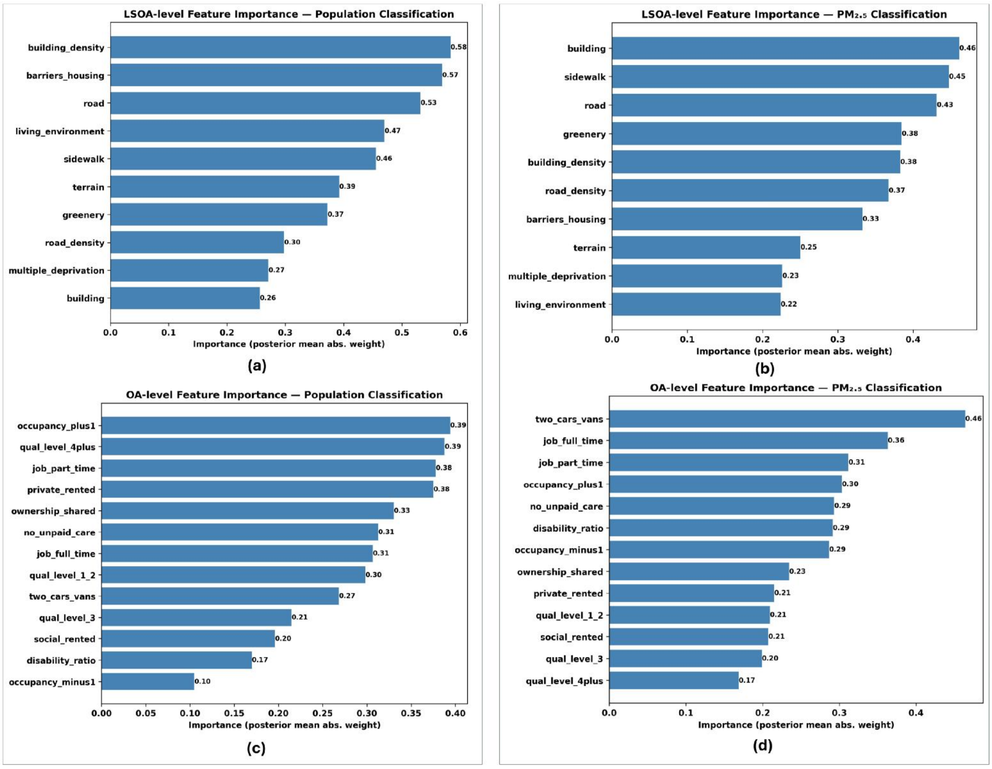
Appendix Table B1. Spatial block cross-validation performance of SR-BHGNN compared to the default random-split performance in Table 2.

Split type	Acc	Prec	Rec	F1	R^2	RMSE	MAPE (%)
Random 70/30 (main text)	0.85	0.92	0.72	0.81	0.58	7234.19	39.78
Spatial block CV (mean over 5 folds)	0.80	0.88	0.69	0.77	0.51	8050.31	43.27

Appendix C. Feature-importance diagnostics at two hierarchical levels

To give readers an intuitive sense of what drives SR-BHGNN's predictions, we compute *global* feature importance separately for OA-level and LSOA-level attributes. For each node type, we extract the posterior mean coefficients from the corresponding Bayesian linear layer of the fitted SR-BHGNN and rank features by the absolute magnitude of these coefficients, averaging over posterior samples. This yields a model-native, scale-specific measure of the contribution of each predictor to the latent representation used for classification. We emphasise that these values are conditional on the fitted model and evaluation data; different splits or model specifications could lead to different rankings. Appendix Figure C1 present the top features for the population-classification and PM $_{2.5}$-classification tasks. Panels (a) and (b) show LSOA-level predictors; panels (c) and (d) show OA-level predictors.

For the population classification task, at the LSOA level, SR-BHGNN relies most strongly on built-form and deprivation indicators (building density, barriers to housing, road and sidewalk coverage, terrain, greenery), which provide the broader structural and socio-economic context within which small populations sit. At the OA level, the most influential features shift to household and tenure characteristics (ownership type, occupancy rating, car ownership, qualification levels, part- vs. full-time work), capturing fine-grained differences between small areas. This pattern suggests that SR-BHGNN draws context from the upper scale while utilising detailed household attributes to refine predictions at the lower scale.



Appendix Figure C1. Global feature importance derived from SR-BHGNN posterior coefficients. (a) LSOA-level features – population classification. (b) LSOA-level features – PM$_{2.5}$ classification. (c) OA-level features – population classification. (d) OA-level features – PM$_{2.5}$ classification. Higher values indicate stronger marginal contribution of the feature to the predictive score at the corresponding hierarchical level.

For PM $_{2.5}$ classification, a similar but not identical pattern appears for air-pollution exposure. At the LSOA level, built-form and transport indicators (building, sidewalks, roads, greenery) remain dominant, reflecting their link to emission sources and dispersion environments. However, deprivation variables are less influential here than for the population. At the OA level, mobility and household indicators, such as the number of cars/vans, full-time/part-time employment, and occupancy ratings, become more important, suggesting that local activity and mobility capacity influence fine-scale PM $_{2.5}$ variation captured by the model.

These diagnostics indicate that SR-BHGNN integrates various types of information at each scale and for each task: broad structural and environmental contexts for LSOAs, and finer socio-demographic or mobility patterns for OAs. Such a diagnostic step provides a transparent, scale-specific picture of the model's behaviour.