Classifying urban areas into residential, non-residential and mixed-use proportions using building footprints and geospatial models

Wole Ademola Adewole

adewoledemy@gmail.com

University of Southampton https://orcid.org/0000-0002-7538-9781

Ortis Yankey

University of Southampton https://orcid.org/0000-0002-0808-884X

Edson Utazi

University of Southampton https://orcid.org/0000-0002-0534-5310

Chris Lloyd

University of Southampton https://orcid.org/0000-0001-7435-8230

Samantha Cockings

University of Southampton https://orcid.org/0000-0003-3333-4376

Andrew J Tatem

University of Southampton https://orcid.org/0000-0002-7270-941X

Research Article

Keywords: Urban functional classification, building footprint data, proportional modelling, settlement analysis, building use, Lagos

Posted Date: February 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8773396/v1

Additional Declarations: The authors declare no competing interests.

1 Introduction

Urban settlements are complex and dynamic entities shaped by socioeconomic, demographic, and spatial interactions (Berry & Neils, 2015; Wekesa et al., 2011; Yiannakou et al., 2017). Their form and function influence how resources are distributed, how economies operate, and how risks are managed, making their accurate characterisation essential for sustainable development and urban resilience (Laros & Jones, 2014; Parnell & Robinson, 2012). Within these systems, buildings constitute a core representational layer of settlement structure, providing observable proxies for population distribution (Boo et al., 2022), land-use organisation (Barr et al., 2004), infrastructure planning (Schiller, 2007), energy demand (Amasyali & El-Gohary, 2018), and environmental sustainability (Abolore, 2012). Advances in global building footprint datasets now enable the mapping of urban form at unprecedented scales (Chamberlain et al., 2024; Sirko et al., 2021), yet footprints alone capture only the extent of settlement, not its function. Classifying settlements into functional categories such as residential, non-residential, and mixed-use, offers deeper insight into urban dynamics by linking spatial form to socioeconomic processes. Such classifications are critical for applications ranging from population disaggregation and service accessibility analysis to transport modelling, economic geography, and disaster risk assessment (Huang et al., 2022; Kang et al., 2018; Stevens et al., 2015).

Traditionally, information on urban settlements has been obtained through surveying and cartographic mapping, which, although highly precise, are expensive, labour-intensive, and challenging to update at scale (Bandam et al., 2022; Li et al., 2021). Government agencies in high-income countries often complement these datasets with administrative property records, enabling detailed classification of buildings into residential, non-residential, and mixed-use categories that support planning, taxation, address systems, and service provision (Biljecki et al., 2021; Hurley et al., 2018; Singh et al., 2022). However, such records are rarely available, up-to-date, or accessible across much of the Global South, leaving critical gaps in the functional characterisation of settlements, and these gaps are identified as central to advancing the SDGs and the New Urban Agenda by the United Nations (Caprotti et al., 2017; UN-GGIM, 2018; United Nations, 2014). In response, Volunteered Geographic Information (VGI) platforms such as OpenStreetMap (OSM) have emerged as alternative sources of building footprints and semantic attributes, enabling functional inference beyond morphology alone (Jokar Arsanjani et al., 2015). Yet despite its global reach, OSM coverage and attribute completeness remain uneven, particularly in LMICs, limiting its reliability for robust urban analytics in precisely those settings where improved building-level information is most urgently needed (Zhou et al., 2022).

Recent advances in earth observation and machine learning have expanded the methodological landscape for classifying urban settlements into functional categories, though each approach presents distinct strengths and limitations. One family of methods combines street-view or façade imagery with convolutional neural networks (CNNs), enabling fine-grained, instance-level classification from visual cues such as signage or storefronts; however, performance is constrained by the sparse and uneven availability of street-view data in many LMICs and by the difficulty of scaling per-building predictions across entire cities (Kang et al., 2018). A second group relies on very-high-resolution satellite imagery and deep learning architectures. An example includes using Mask R-CNN to infer building function from roof structure and texture, as demonstrated in the Urban Building Classification (UBC) dataset (Huang et al., 2022), though the weak relationship between roof appearance and functional use, especially in dense

environments with substantial occlusion, limits accuracy. Complementing these vision-based techniques, attribute-driven classifiers and spatial models use covariates such as footprint morphology, road accessibility, points of interest, and population proxies to estimate functional patterns at broader spatial scales. These methods, including Random Forests, support vector machines, and hierarchical Bayesian models, have proven particularly valuable in data-scarce contexts, producing robust grid-level estimates even where building-level semantic data are incomplete (Stevens et al., 2015). Notable applications include pattern-based settlement classification in Afghanistan using morphological metrics (Jochem et al., 2018), residential–non-residential modelling in LMIC cities using geospatial predictors (Lloyd et al., 2020), and multiscale footprint-derived settlement characterisation across heterogeneous urban fabrics (Jochem et al., 2021). More recently, deep learning models using OSM-derived ground truth have enabled large-scale land-use classification across East African cities (Rosier et al., 2025). Additionally, Oostwegel et al. (2025) derived building footprint functions by combining available building attributes with contextual information from OSM and the Global Human Settlement Layer. Together, these approaches underscore the rapid methodological progression in functional settlement mapping, while highlighting the persistent challenge of developing scalable, transferable models suited to regions with limited administrative data.

Much of the existing literature treats urban function as a discrete property, assigning spatial units (e.g., grid cells, buildings, or neighbourhoods) to a single dominant category such as residential or commercial. While this simplifies classification, it fails to reflect the substantial functional mixing that characterises many contemporary cities, especially in rapidly growing urban environments where mixed-use buildings and parcels are commonplace. A single-class assignment at 100 m resolution, for example, implicitly assumes homogeneity within each grid cell and obscures the heterogeneity of buildings that may simultaneously accommodate retail, services, and housing. This limitation is particularly pronounced in urban areas in LMICs where mixed-use configurations are prevalent and administrative records are often incomplete or outdated. To address this gap, we introduce a proportional modelling framework that estimates the relative shares of residential, non-residential, and mixed-use activity within each grid cell, leveraging building footprints, a property administrative record dataset (Lagos Building Sample Point - LBSP) and a suite of geospatial predictors. By treating urban function as a compositional mixture rather than a discrete label, the approach captures fine-grained heterogeneity and generates semantically richer datasets that are better suited for downstream applications such as population disaggregation, transport modelling, and urban infrastructure planning.

2 Methods

Effective classification of urban areas into functional classes such as residential, non-residential and mixed-use requires a robust methodological framework that integrates diverse data sources while accounting for spatial heterogeneity. To achieve this, we leverage a combination of geospatial covariates, and implement using machine learning, and geostatistical modelling. The use of covariates, such as building morphology, population density, and socioeconomic indicators, allows for a comprehensive characterisation of urban environments, capturing key spatial patterns that influence building functionality. To ensure the representativeness of each building class per unit area of prediction, a gridded approach of 1 x 1km spatial resolution was adopted. This ensures that each building class observation, though sparsely distributed in the LBSP dataset, which is used as the model training samples, are proportionally represented at the grid cell level. In addition, this ensures consistency in spatial

representation, facilitates integration with auxiliary datasets, and enhances model generalizability across different urban typologies. By independently implementing our urban area classification method in a machine learning and Bayesian hierarchical model, we can compare its performance in both model approaches and understand if any bias can be introduced due to the selected model. This methodological integration enables a nuanced and scalable approach to urban classification, offering insights applicable to diverse urban settings. The rest of this methods section is organised into four parts. The first part introduces the study area, and the LBSP dataset. The second part discusses the processing and selection of the covariate datasets. In the third part, we provide an overview of the modelling framework used for urban area classification, focusing on machine learning and a geostatistical approach**.** We bring this together in the fourth part by examining the performance of the models.

2.1 Lagos building sample point dataset (LBSP)

The LBSP dataset, collected in 2018 by the Lagos state government, aimed to survey all buildings within Lagos state, covering residential, commercial, and mixed-use buildings, to create a comprehensive property record database (Lagos Land Use Agency, 2018). The survey methodology involved both physical observations, capturing visible building characteristics, and interviews of residents to gather additional information not immediately observable. Field officers used real-time mapping software (GIS Cloud, https://www.giscloud.com/) for navigation and data capture, standing on adjoining access roads to observe buildings they couldn't access due to residents out-of-home or other reasons. Revisits were not conducted for inaccessible buildings, and due to shifted government priorities, the survey was not completed. Nonetheless, the dataset provided over 180,000 building points with attribute information (Figure 1 and Supplementary Table A.1), used as a training sample dataset in this research. As one of Nigeria's and Africa's largest and fastest-growing cities, Lagos, with its complex, heterogeneous, and multifaceted settlement characteristics (Oduwaye, 2008), presents an ideal case study, embodying Nigeria's role as a commercial, tourism, industrial, and social hub for West Africa.

The building functions used in this study follow the classifications defined within the LBSP dataset.

Residential: This consists of where people live, mainly in houses and apartments.

Non-residential: This consists of buildings for sole commercial purposes, including supermarkets, hotels, office buildings, etc.

Mixed-use: This consists of buildings which has a combination of residential functions (where people live), and non-residential functions

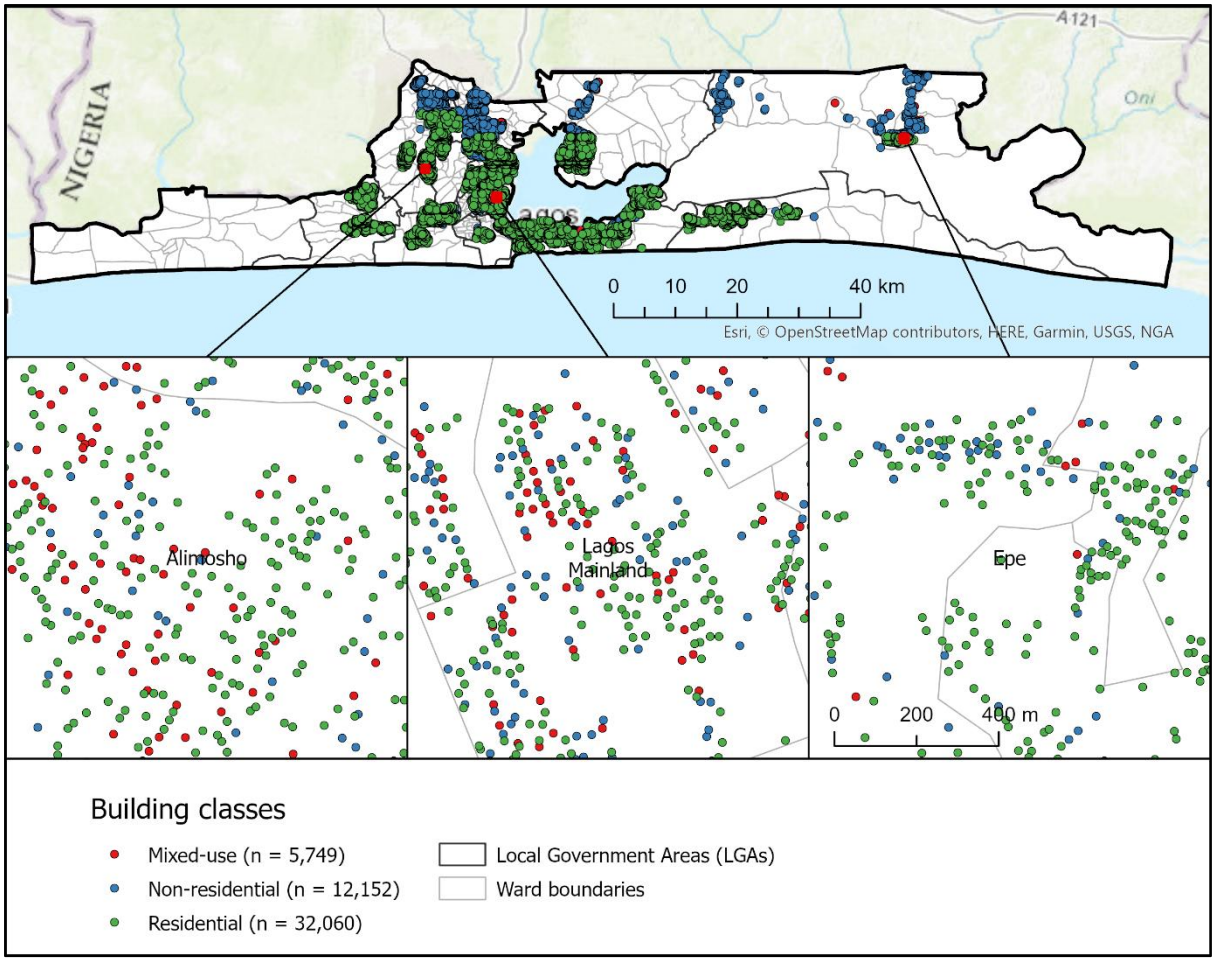

Figure 1: Spatial distribution of Lagos Building Sample Points (LBSP) showing residential, non-residential and mixed-use building points with an insert map of specific areas

2.2 Dataset Processing and Covariate Selection

We produced 68 gridded covariate datasets. These include the standard datasets used in the study of population distribution (Nieves et al., 2017), land use information, socioeconomic variables, and points of interest (Supplementary Table A.2). We first constrained these datasets to the extent of the study area and aggregated to 1 x 1 km resolution to align with the spatial scale the LBSP observation dataset was aggregated to. We then selected model covariates by assessing the multicollinearity among the covariates. We employed the Variance Inflation Factor (VIF), which helps quantify the degree of multicollinearity by assessing how much the variance of a regression coefficient is inflated due to correlations among the predictors. Covariates with VIF values exceeding a predefined threshold (commonly VIF > 5) were removed, aligning with established literature on handling highly correlated variables (Kalnins & Praitis Hill, 2025). This was performed for the proportion of residential and non-residential building classes, which enabled us to select 19 covariates. The selected covariates were standardised using a z-score.

A total of 49,961 observations in LBSP dataset had building class information, of which 32,060 were residential, 12,152 were non-residential and 5,749 were mixed-use buildings (Figure 1). The total count of each building class was aggregated to a 1 x 1 km grid cell. Per 1 x 1 km grid cell, the count of residential (B_{res}), non-residential (B_{nonres}), and mixed-use (B_{mixed}) buildings is equivalent to the total count of buildings per grid cell (B_{total}) as shown in Equation (1) and

visually illustrated in Figure 2. Therefore, we can derive the proportion of each building class per grid cell as shown in Equation (2). Because modelling residential, non-residential, and mixed-use classes independently would not guarantee that their predicted proportions sum to one within each grid cell, we instead modelled only the residential and non-residential components directly and derived the mixed-use proportion using Equation (3). This approach ensures internal coherence across the three functional categories, yielding proportional estimates that are mathematically constrained to sum to one. While the limited number of mixed-use observations in the LBSP dataset further reinforced the rationale for this strategy, the primary motivation was to maintain a consistent and interpretable compositional structure in the final predictions.

$$Btotal = Bres + Bnonres + Bmixed \tag{1}$$

$$if\ 1 = \left(\frac{Bres}{Btotal}\right) + \left(\frac{Bnonres}{Btotal}\right) + \left(\frac{Bmixed}{Btotal}\right) \tag{2}$$

$$\frac{Bmixed}{Btotal} = 1 - \left(\frac{Bres}{Btotal} + \frac{Bnonres}{Btotal}\right) \tag{3}$$

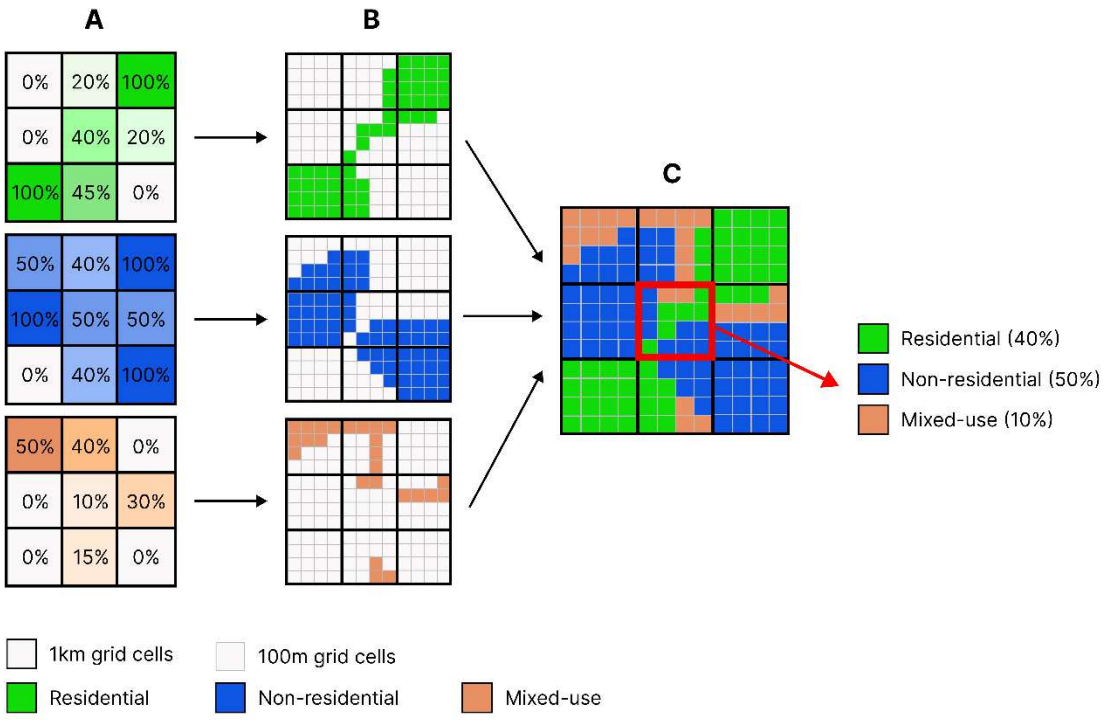

Figure 2: Illustrative diagram of the proportional modelling framework for computing residential, non-residential and mixed-use shares of functional uses per 1km grid cell. A) shows grid cells at 1km grid cell with proportion of functional classes ratio when separated; B) shows a disaggregation of 1km grid cells into sub-grid to highlight the distribution of functional classes when separated, and; C) shows the combination of the functional classes at 1km with sub-grid cell distributions and their respective proportions

2.3 Urban Area Classification Model

We applied two modelling approaches: machine learning and geostatistical modelling. For the machine learning approach, a Random Forest model is used, while for the geostatistical

approach, a Bayesian Hierarchical model is adopted using an Integrated Nested Laplace Approximation (INLA) model combined with a Stochastic Partial Differential Equation (SPDE).

2.3.1 Random Forest

Random Forest (RF) is an ensemble machine learning algorithm capable of handling complex non-linear relationships and interactions among predictors. In addition to the selected covariates, we calculated spatial lags for the proportion of residential and non-residential buildings within each grid cell to incorporate spatial dependencies into the RF model. The spatial lags were derived using distance-based spatial weights, computed using the geographical coordinates (x, y) of the grid cell centroids (Georganos & Kalogirou, 2022). A distance threshold of 0.01 and 0.05 was selected for constructing spatial weights for non-residential and residential models, respectively, and these weights were row-standardised to ensure comparability across grid cells with differing numbers of neighbours. The spatial lag was then computed as the weighted average of neighbouring values. The spatial lag was added as one of the covariates used in the RF model.

The RF model construction involved growing an ensemble of decision trees using bootstrap samples of the training data, where a random subset of predictors was evaluated at each tree split to minimise correlation among trees and enhance model diversity. Hyperparameter tuning was conducted through a systemic grid search over a range of values for the number of predictors considered at each split *(mtry = 23)* and the number of trees *(ntree = 90)*.

2.3.2 Bayesian Hierarchical

A Bayesian Hierarchical (BH) modelling approach was also employed to predict the proportion of building classes at 1 x 1 km spatial resolution, using an INLA – SPDE framework. The model was trained on a dataset containing spatial covariates and building class proportion as the outcome variable. The geographical coordinates (x, y) of the grid cell centroids define both estimation and prediction points. A spatial mesh was constructed using the boundary information of Lagos and the geographic distribution of the centroid of the grid cells to capture the spatial dependency structure within the study area. The mesh was designed with adaptive resolutions, using the parameters: max edge (0.2, 0.5), offset (0.05, 0.2), and cutoff (0.01), ensuring finer granularity near densely sampled areas and coarser resolution elsewhere. The fitted model is detailed in Equation (4) and (5), where $Y(\boldsymbol{s})$ denotes the outcome variable at spatial location $\boldsymbol{s}$ (denoting the longitude and latitude coordinates) which is assumed to follow a beta distribution with mean $\mu(\boldsymbol{s})$ and dispersion parameter ϕ. Furthermore, $g(.)$ is the link function (logit link), while $\mathbf{x}(\boldsymbol{s})$ is a vector of geospatial covariates and $\boldsymbol{\beta}$, the corresponding regression coefficients. $\omega(\boldsymbol{s})$ and $\epsilon(\boldsymbol{s})$ are spatial identical and independently distributed random effects, respectively.

$$Y(\boldsymbol{s}) \sim Beta(\mu(\boldsymbol{s}), \phi) \tag{4}$$

$$g\big(\mu(\boldsymbol{s})\big) = \mathbf{x}^T(\boldsymbol{s})\boldsymbol{\beta} + \omega(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \tag{5}$$

The SPDE framework was applied to construct a spatial Gaussian Markov Random Field (GMRF) over the mesh, modelling spatial correlations in the data. A matrix was created to map the

GMRF values from the mesh nodes to the observation and prediction locations. Also, the spatial model incorporated a Matern covariance structure with priors on the range and variance parameters to ensure appropriate regularisation. The hierarchical model included a fixed-effect component for the covariates, a spatial random effect for the SPDE, and an independent random effect accounting for unstructured residual variability. A beta regression likelihood was applied to accommodate the proportional nature of the response variable. Model fitting was performed using INLA, with hyperparameters estimated through penalised complexity priors. The fitted model was used to generate spatial predictions and summarised using their mean. Additionally, the 95% credible intervals for the predictions were derived by extracting the 2.5th and 97.5th percentiles, providing bounds for the uncertainty in the estimates.

2.4 Model fit and validation

RF and BH models independently predicted building classes using geospatial covariates. Both models were validated using k-fold cross-validation, with model performance evaluated through in-sample and out-of-sample predictions based on metrics presented in Table 1. For the BH model, predictive performance was assessed using metrics (Table 1) computed for in-sample and out-of-sample predictions alongside deviance information criterion (DIC), Watanabe-Akaike information criterion (WAIC), and conditional predictive ordinates (CPO) for model comparison and validation. Both models utilised a 10-fold cross-validation framework to ensure robust hyperparameter optimisation and mitigate overfitting (Malakouti et al., 2023).

Table 1: Statistical metrics calculated to evaluate the performance of both Random Forest and Bayesian Hierarchical models.

Metrics	**Description**
Mean Square Error (MSE)	Measures average squared differences between observed and predicted values.
Root Mean Square Error (RMSE)	Quantifies prediction error by averaging squared residuals and taking the square root.
Pearson Correlation (r)	Quantifies the linear relationship between two variables, ranging from -1 to 1.
Pseudo R-Squared (R^2)	Measures the proportion of variance in the dependent variable explained by the independent variables.

Predictions from the RF and BH models were compared with observation values using density plots to assess the consistency and reliability of the models.

3 Results

Predictions were generated at a 1km resolution, with each grid cell assigned a value representing the proportion of residential, non-residential, and mixed-use buildings, ranging from 0 to 100. This resolution was chosen to ensure that each observation cell contains a sufficient number of buildings to produce stable and representative functional proportions. The resulting data output is an estimated proportion of building classes (residential, non-residential, and mixed-use) for each 1km grid cell in Lagos State, as defined by its administrative boundary.

3.1 Statistical comparison of random forest and Bayesian hierarchical models

Using geospatial covariates, the RF and BH models exhibited distinct advantages in predicting the proportions of residential and non-residential building classes across Lagos, Nigeria. The RF model demonstrated strong predictive capabilities, effectively capturing complex non-linear relationships among covariates (Table 2). This translated into robust performance metrics for both in-sample and out-of-sample predictions. For residential building proportions, the RF model achieved a Root Mean Square Error (RMSE) of 0.09 and 0.19 for in-sample and out-of-sample predictions, respectively, with high correlation coefficients of 0.98 and 0.85, and pseudo-R^2 values of 0.96 and 0.73. Non-residential predictions exhibited a similar trend, with an in-sample RMSE of 0.09 and an out-of-sample RMSE of 0.22, alongside pseudo-R^2 values of 0.94 and 0.52, and correlation coefficients of 0.97 and 0.72.

Table 2: Goodness of fit metrics for Random Forest and Bayesian Hierarchical predictions

Models	**Classes**	**Predictions**	**(MAE)**	**(RMSE)**	**Pearson (r)**	**R^2**
Random Forest	Residential	In-sample	0.06	0.09	0.98	0.96
		Out-of-sample	0.13	0.19	0.85	0.73
	Non-residential	In-sample	0.06	0.09	0.97	0.94
		Out-of-sample	0.15	0.22	0.72	0.52
Bayesian Hierarchical	Residential	In-sample	0.1	0.13	0.96	0.92
		Out-of-sample	0.14	0.21	0.84	0.70
	Non-residential	In-sample	0.02	0.09	0.83	0.64
		Out-of-sample	0.12	0.22	0.69	0.51

* MAE: Mean Absolute Error, RMSE: Root Mean Squared Error

The Bayesian Hierarchical (BH) model, leveraging the spatial structure of the data through a spatial random effect explicitly modelled using a Matern covariance function, offered comparable predictive performance while providing additional insights into uncertainty (Supplementary Figure A.1). For residential building proportions, the BH model achieved an in-sample RMSE of 0.13 and an out-of-sample RMSE of 0.21, with pseudo-R^2 values of 0.92 and 0.70, and correlation coefficients of 0.96 and 0.84, respectively. Predictions for non-residential buildings were consistent, with an RMSE of 0.09 for both in-sample and 0.22 for out-of-sample predictions, pseudo-R^2 values of 0.64 and 0.51, and correlation coefficients of 0.83 and 0.69. Exploring the prediction accuracy at Local Government Area (LGA) level, a scatter plot of the predicted and observed values is plotted for each model output (Supplementary Figures A.2 and A.3), and this shows that the BH has higher correlation coefficients compared with RF for the residential and non-residential predictions.

We further evaluated the predictive performance of the BH and RF models using density plots (Figure 3), evaluated against the LBSP observational dataset. The BH model shows closer alignment with observed distributions for residential, non-residential, and derived mixed-use proportions, reflecting its ability to capture broader spatial variability and uncertainty. By contrast, the RF model constrains predictions into a narrower, near-normal distribution, suggesting a degree of oversmoothing and reduced sensitivity to the full range of functional heterogeneity. While RF achieves higher point-level accuracy, the BH model provides more realistic distributional behaviour and coherence with empirical patterns. Together, the two models reveal complementary strengths. RF offers strong predictive power, and BH delivers improved representation of real-world variability, underscoring the methodological value of deterministic and probabilistic approaches in urban function modelling.

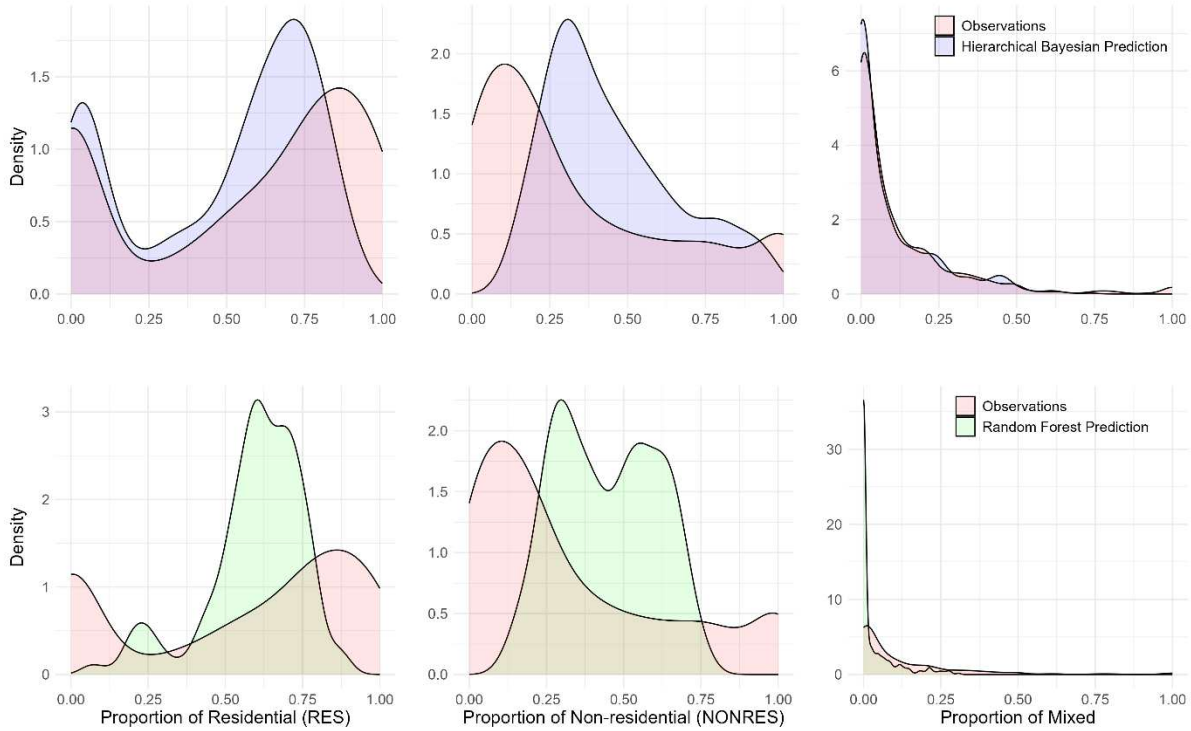

Figure 3: Density plot for the predicted proportions of residential (RES), non-residential (NONRES) and mixed-use building classes from Random Forest and Bayesian Hierarchical models compared to the true proportions per 1 x 1 km grid cell.

3.1.1 Residential and non-residential classes

The spatial distribution of residential class predictions for both RF and BH models exhibits a clear clustering in central and southern Lagos, with grid cells exceeding 0.5 in proportion (Figure 4). This includes major parts of Ojo, Alimosho, Oshodi-Isolo, Mushin, Shomolu, and Lagos Mainland LGA, extending eastward into Ibeju Lekki and Eti-Osa. These areas represent the dense residential fabric characteristic of urban and peri-urban settlements. However, differences emerge in the predicted proportions, with the BH model generally yielding lower values than the RF model (mean: 0.378 vs. 0.50; median: 0.41 vs. 0.53). Notably, the BH model indicates pockets of lower residential class (< 0.5) within central Lagos, particularly in Mushin, Ikeja, Surulere, Amuwo Odofin, Lagos Mainland, Lagos Island, Apapa, and Ajeromi Ifelodun.

Conversely, the non-residential class predictions demonstrate an inverse pattern, emphasising economic and industrial hubs. Higher proportions (> 0.5) are observed in parts of central Lagos, notably in Ikeja, Apapa, Lagos Island, Agege, Ifako Ijaiye, and parts of Eti-Osa and Alimosho, aligning with commercial and infrastructural corridors (Figure 4). Here, the BH model predicts systematically higher non-residential proportions than the RF model (mean: 0.44 vs. 0.32; median: 0.41 vs. 0.26). In suburban areas like Ikorodu, the BH model estimates a greater extent of non-residential development (> 0.5 in 56.5% of grid cells) compared to the RF model (19.6%). However, in select LGAs, such as Ikeja, the RF model predicts greater non-residential dominance (95.7% vs. 80.8% of grid cells > 0.5 in RF and BH models, respectively).

3.1.2 Mixed-use class

The mixed-use class proportions, derived post-modelling, reveal distinct spatial trends across the two models. The RF model predominantly assigns higher mixed-use proportions to

suburban areas such as Ikorodu, Badagry, Ibeju Lekki, and parts of Epe, whereas the BH model exhibits spatial clustering of high mixed-use values in central Lagos, particularly in Ikeja, Alimosho, Lagos Island, Kosofe, Agege, Amuwo Odofin, and Ikorodu (Figure 5). Notably, 25.6% and 48.5% of grid cells in Ikorodu and Kosofe exceed the 0.4 threshold for mixed-use in the BH model, reflecting the functional integration of residential and non-residential activities in these urban centres. Comparative analysis indicates that the RF model generally predicts higher mixed-use proportions than the BH model (maximum: 0.73 vs. 0.67; mean: 0.20 vs. 0.18; median: 0.12 vs. 0.14). This suggests that the RF model is more sensitive to mixed-use patterns or potentially overestimates their extent compared to the BH model.

Overall, the classification results underscore the heterogeneity of urban area functional use in Lagos, capturing the dominant residential clustering in the urban core, the concentration of commercial and industrial activities in key economic zones, and the variation in mixed-use development across different spatial scales. While both models reveal similar spatial trends, their differences highlight the influence of model assumptions and parameterisation. The RF model consistently predicts higher residential and mixed-use proportions, whereas the BH model emphasises non-residential clustering in economic hubs. These findings provide critical insights for urban planning, land-use optimisation, and sustainable city development in Lagos.

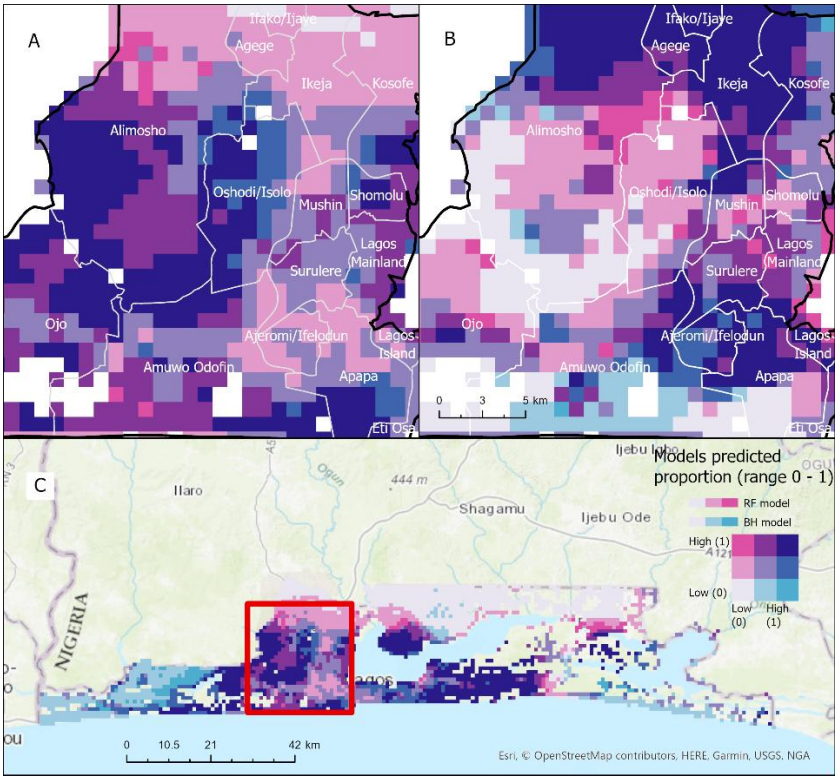

Figure 4: Bivariate map of the predicted proportion of classes (range 0 – 1) for Random Forest (RF) and Bayesian Hierarchical (BH) models per 1 x 1km grid cell across Lagos; (A) Proportion of residential class; (B) Proportions of non-residential class; (C) Proportions of residential class showing the entirety of Lagos.

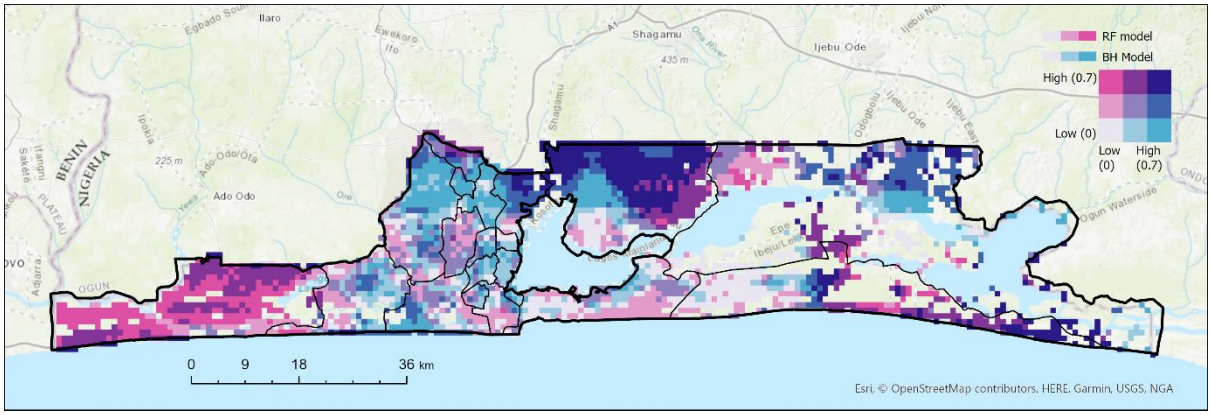

Figure 5: Bivariate map of the derived mixed-use proportion (range 0 – 0.7) for Random Forest (RF) and Bayesian Hierarchical (BH) models per 1 x 1 km grid cell across Lagos

Figure 6 illustrates an example of this integration, where the predicted proportions are combined with Google Building Footprints (version 2) dataset (Google Research, 2024) by a direct multiplication (*proportion of settlement classes x building footprint count or area*). The resulting visualisation shows the spatial heterogeneity of building counts across Lagos. This clearly shows the variability of the RF and BH model results but also highlights density of functional classes across LGAs like Alimosho. This approach enhances the interpretability of the model outputs and allows for integration with other geospatial data products, creating a dataset that can support targeted interventions. The utility of these data products extends across multiple domains. Urban planners can leverage them to assess housing density and infrastructure needs (Schiller, 2007), while policymakers can identify regions of mixed-use development for strategic economic planning (Wekesa et al., 2011). Additionally, disaster response teams can use the spatial distribution of building types to estimate population exposure and vulnerability in emergency scenarios.

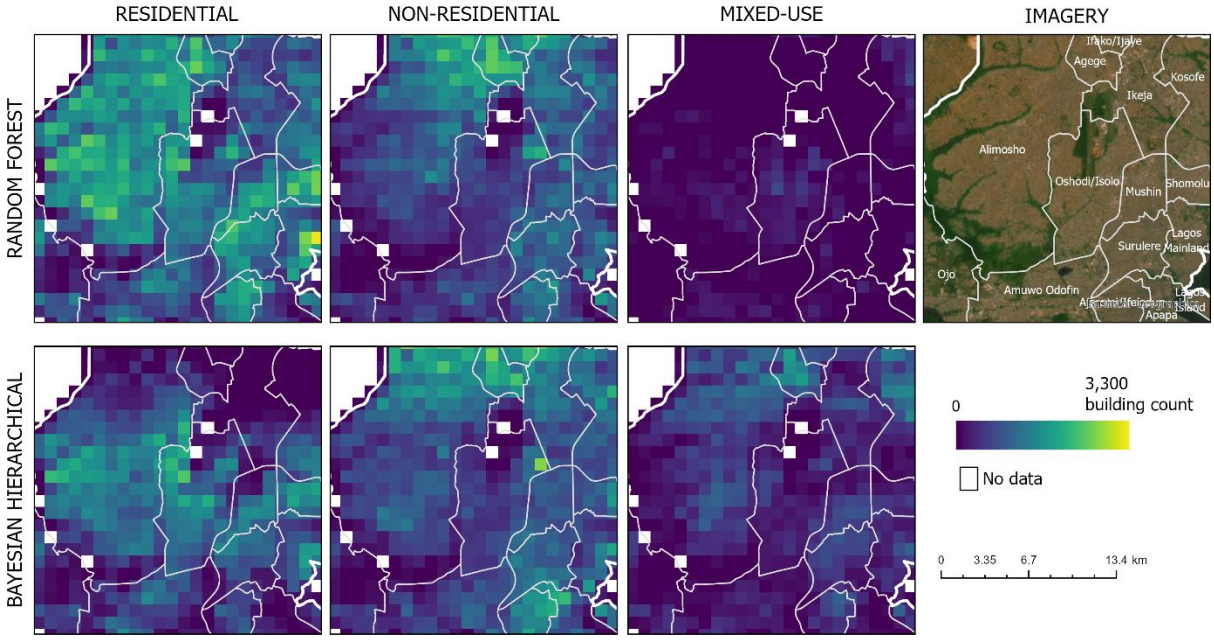

Figure 6: The total count of residential, non-residential and mixed buildings in Alimosho, Mushin, Oshodi Isolo and other areas in Lagos. This is produced using the formula (predicted proportion of settlement classes x count of Google building version 2) from the Random Forest and Bayesian Hierarchical models at 1 x 1 km.

4 Discussion

In this research, we introduced a novel approach to classify 1 x 1km grid cells into proportions of residential, non-residential and mixed-use classes and implemented this across Lagos. Our result revealed a high presence of residential buildings (> 0.5) across most parts of Lagos, whereas non-residential presence was higher in the centre but also extending towards the eastern areas of the state (Figure 4). Mixed-use presence, which was derived post-modelling, was higher in the northern parts of the state. While very high resolution satellite images, LiDAR datasets, and address classification databases/systems have been adopted for building classification at unit level in some studies (Bandam et al., 2022; Huang et al., 2022; Lu et al., 2014; Smith & Crooks, 2010), our approach focuses more on compositional predictions at the grid cell level with additional insight on the share of uses within each grid. The method allows for universal spatial consistency which is advantageous for ensuring global or regional comparability, one of the benefits of grid-based systems (Lloyd et al., 2020). In addition, being able to predict proportions of functional use classes per grid cell makes it robust enough to expand to multi-classes and be implemented in different models. It is not only difficult to determine unit level building function from just visual features of satellite imagery tiles, especially in urban areas (Huang et al., 2022), as well as computationally expensive (Bandam et al., 2022). Our ability to predict at a coarser level, which reduces the computational dependency but still provides insight at sub-grid-cell levels.

Though the prediction of the RF and BH models had a generally similar spatial pattern, the grid cell level proportions for each of the models varied. The RF model exhibited higher prediction values overall for residential and non-residential classes with correlation coefficients of 0.85 and 0.72 for out-of-sample, respectively, as shown in Table 2. Meanwhile, the BH model demonstrated slightly lower prediction with correlation coefficients of 0.84 and 0.69 for residential and non-residential, respectively. Though from the RMSE, R^2, and correlation values as detailed in the performance metrics (Table 2), the RF and BH models look very similar in their predictive abilities, the density plot in Figure 3 and LGA level scatterplot (Supplementary Figure A.2 and A.3) reveals that the BH performs better as the predictions are the closest to the LBSP observation values for the three classes. This finding underscores the importance of probabilistic approaches in urban research, where building uses are not always mutually exclusive and may exhibit significant spatial variability. In addition, a distinguishing feature of the BH model is its ability to generate 95% credible intervals and coefficients of variation (Lu et al., 2012). This allows for a more refined interpretation of prediction reliability across space. This capability is particularly advantageous in urban contexts like Lagos, where spatial heterogeneity and the scarcity of comprehensive ground-truth datasets present significant challenges.

In this research, we reveal clear functional clustering within Lagos's urban fabric. Residential buildings are strongly concentrated in central and peri-urban districts such as Alimosho, Ojo, Oshodi-Isolo, and Mushin, reflecting long-recognised demographic pressures of rapid migration and housing shortages (Oluwoye, 2008). This residential dominance echoes earlier accounts of Lagos's polycentric yet heavily residential urban morphology (Abiodun, 1997) and provides quantitative confirmation of patterns previously documented through census and qualitative studies. In contrast, non-residential clusters are concentrated in Ikeja, Lagos Island, Apapa, and Eti-Osa, localities historically associated with governance, commerce, and industry (Gandy, 2006; Oduwaye, 2008). The higher non-residential proportions identified by the Bayesian Hierarchical model are consistent with Lagos's established economic geography,

including Ikeja's administrative role and Apapa's prominence as Nigeria's primary port. These findings align with studies that describe the city's transition from a largely monocentric to increasingly polycentric urban economy, characterised by highly localised employment centres (Abiodun, 1997). By resolving these patterns at 1 km^2 resolution, this study offers fine-grained, data-driven evidence of long-recognised economic and spatial structures.

Mixed-use development emerges as another defining feature, particularly in dense neighbourhoods such as Alimosho, Kosofe, and parts of Lagos Island. These patterns illustrate the integration of residential and commercial activities typical of African urbanism, where formal and informal economies frequently coexist within the same built environment (Simone, 2004). Such mixed-use development has been highlighted as an adaptive response to inadequate planning, land scarcity, and informality (Watson, 2009). The Bayesian model's identification of central hotspots for mixed-use further resonates with ethnographic accounts of multifunctional streetscapes in Lagos (Fourchard, 2011). Collectively, these findings demonstrate the functional heterogeneity of Lagos and underscore the value of geospatial modelling in capturing the dynamics of one of Africa's most complex urban systems.

The LBSP dataset used in this research is a rich dataset providing labels of residential, non-residential, and mixed-use buildings. This dataset was used in training RF and BH models in this research, but it also has its limitations. First, the original intent of the data collection was to build a property record for Lagos; as such, the spatial distribution of the collected data was not stratified to cover representative parts of the whole state e.g. there is no LBSP in Badagry LGA (Lagos Land Use Agency, 2018). In addition, the LBSP is a subset of all buildings within the 1 x 1km grid cell extent. There is thus a possibility that the representativeness of the proportion of the building classes is skewed towards some classes. As such, though the methodology developed in this research is independent of this limitation, using Equation (1) and (2) as a relationship for our test in Lagos might have skewed the proportion of classes labelled. A solution would be improving and updating the LBSP or using a ground survey dataset that has an equitable distribution of building class information for all buildings per grid cell, to ensure more accurate and reliable predictions. In theory, better completeness and accuracy of data in training allows higher quality predictions to be made (Lloyd et al., 2020).

In this research, we have advanced the study of classifying urban areas into functional uses, but there are still opportunities to extend this research. Areas of future studies can look into increasing the spatial resolution of outcome predictions from 1km to 100m or finer. The spatial resolution of the outcome predictions is driven by the cluster of training samples that represent all classes per grid cell. Also, the spatial distribution of predicted proportions of residential, non-residential and mixed-use patterns can be compared with existing masterplans for Lagos. Our research has piloted the developed methodology in Lagos as a case study, but future attempts can extend the application to other contexts and regions to validate the adaptability of this approach across settings. The methodology developed depends on the availability of ground truth training samples of building use information to fit the model. This provides an opportunity for further refinement of the method to be adaptable with limited or no building use information for training and rather depend on secondary proxy data.

5 Conclusions

The inherent complexity of urban systems, marked by informal settlements, mixed-use developments, and rapid urbanisation in cities like Lagos, requires innovative approaches. In

this study, we introduce a novel functional-use classification framework applied in BH and RF models. The output characterises buildings at 1 km resolution to overcome key limitations of traditional approaches. The resulting proportional estimates can be paired with any building footprint or settlement layer to derive absolute counts, offering a flexible input for population modelling and urban analytics. The insights from this study have significant implications for urban studies, offering practical tools to improve population estimates, resource allocation, and policy-making in rapidly urbanising regions. Looking ahead, combining such modelling approaches with improved ancillary datasets and participatory data collection presents a promising direction for refining urban functional mapping and strengthening evidence-based planning in rapidly urbanising regions.

6 Reference

Abiodun, J. O. (1997). The challenges of growth and development in metropolitan Lagos. *The urban challenge in Africa: growth and management of its large cities*.

Abolore, A. A. (2012). Comparative study of environmental sustainability in building construction in Nigeria and Malaysia. *Journal of Emerging Trends in Economics and Management Sciences*, *3*(6), 951-961.

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192-1205.

Bandam, A., Busari, E., Syranidou, C., Linssen, J., & Stolten, D. (2022). Classification of Building Types in Germany: A Data-Driven Modeling Approach. *Data*, *7*(4), 45. https://www.mdpi.com/2306-5729/7/4/45

Barr, S. L., Barnsley, M. J., & Steel, A. (2004). On the separability of urban land-use categories in fine spatial scale land-cover data using structural pattern recognition. *Environment and Planning B: Planning and Design*, *31*(3), 397-418.

Berry, B. J., & Neils, E. (2015). Location, size, and shape of cities as influenced by environmental factors: the urban environment writ large. In *The Quality of the Urban Environment* (pp. 255-302). Routledge.

Biljecki, F., Chew, L. Z. X., Milojevic-Dupont, N., & Creutzig, F. (2021). Open government geospatial data on buildings for planning sustainable and resilient cities. *arXiv preprint arXiv:2107.04023*.

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W., & Tatem, A. J. (2022). High-resolution population estimation using household survey data and building footprints. *Nature Communications*, *13*(1), 1330. https://doi.org/10.1038/s41467-022-29094-x

Caprotti, F., Cowley, R., Datta, A., Broto, V. C., Gao, E., Georgeson, L., Herrick, C., Odendaal, N., & Joss, S. (2017). The New Urban Agenda: key opportunities and challenges for policy and practice. *Urban research & practice*, *10*(3), 367-378.

Chamberlain, H. R., Darin, E., Adewole, W. A., Jochem, W. C., Lazar, A. N., & Tatem, A. J. (2024). Building footprint data for countries in Africa: to what extent are existing data products comparable? *Computers, Environment and Urban Systems*, *110*, 102104.

Fourchard, L. (2011). Between world history and state formation: new perspectives on Africa's cities. *The Journal of African History*, *52*(2), 223-248.

Gandy, M. (2006). Planning, anti-planning and the infrastructure crisis facing metropolitan Lagos. *Urban Studies*, *43*(2), 371-396.

Georganos, S., & 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.

Google Research. (2024). *Open buildings: A dataset of building footprints to support social good applications*. https://sites.research.google/open-buildings/#faq

Huang, X., Ren, L., Liu, C., Wang, Y., Yu, H., Schmitt, M., Hänsch, R., Sun, X., Huang, H., & Mayer, H. (2022). Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Hurley, J., Wood, G., & Groenhart, L. (2018). Long run urban analysis using property records: A methodological case study of land use change. *Urban Studies*, *55*(2), 427-442.

Jochem, W. C., Bird, T. J., & Tatem, A. J. (2018). Identifying residential neighbourhood types from settlement points in a machine learning approach. *Computers, Environment and Urban*

Systems, *69*, 104-113. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2018.01.004

Jochem, W. C., Leasure, D. R., Pannell, O., Chamberlain, H. R., Jones, P., & Tatem, A. J. (2021). Classifying settlement types from multi-scale spatial patterns of building footprints. *Environment and Planning B: Urban Analytics and City Science*, *48*(5), 1161-1179. https://doi.org/10.1177/2399808320921208

Jokar Arsanjani, J., Mooney, P., Zipf, A., & Schauss, A. (2015). Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In *OpenStreetMap in GIScience: experiences, research, and applications* (pp. 37-58). Springer.

Kalnins, A., & Praitis Hill, K. (2025). The VIF score. What is it good for? Absolutely nothing. *Organizational research methods*, *28*(1), 58-75.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., & Zhu, X. X. (2018). Building instance classification using street view images. *Isprs Journal of Photogrammetry and Remote Sensing*, *145*, 44-59. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2018.02.006

Lagos Land Use Agency. (2018). *Dataset on property records from Lagos data capturing and enumeration exercise*.

Laros, M., & Jones, F. (2014). The state of African cities 2014: re-imagining sustainable urban transitions.

Li, Z., Xin, Q., Sun, Y., & Cao, M. (2021). A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery. *Remote Sensing*, *13*(18), 3630.

Lloyd, C. T., Sturrock, H. J., Leasure, D. R., Jochem, W. C., Lázár, A. N., & Tatem, A. J. (2020). Using GIS and machine learning to classify residential status of urban buildings in low and middle income settings. *Remote Sensing*, *12*(23), 3847.

Lu, D., Ye, M., & Hill, M. C. (2012). Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification. *Water resources research*, *48*(9).

Lu, Z., Im, J., Rhee, J., & Hodgson, M. (2014). Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landscape and Urban Planning*, *130*, 134-148. https://doi.org/https://doi.org/10.1016/j.landurbplan.2014.07.005

Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2023). The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Cleaner Engineering and Technology*, *15*, 100664.

Nieves, J. J., Stevens, F. R., Gaughan, A. E., Linard, C., Sorichetta, A., Hornby, G., Patel, N. N., & Tatem, A. J. (2017). Examining the correlates and drivers of human population distributions across low-and middle-income countries. *Journal of the Royal Society interface*, *14*(137), 20170401.

Oduwaye, L. (2008). Planning implications of the ethnic structure of residential areas of Metropolitan Lagos. *Asian Social Science*, *4*(8), 129-136.

Oluwoye, J. (2008). An assessment of why the problems of housing shortages persist in developing countries: A case study of Lagos Metropolis, Nigeria.

Oostwegel, L. J. N., Schorlemmer, D., & Guéguen, P. (2025). From Footprints to Functions: A Comprehensive Global and Semantic Building Footprint Dataset. *Scientific Data*, *12*(1), 1699. https://doi.org/10.1038/s41597-025-06132-z

Parnell, S., & Robinson, J. (2012). (Re) theorizing cities from the Global South: Looking beyond neoliberalism. *Urban geography*, *33*(4), 593-617.

Rosier, J. F., Verburg, P. H., & van Vliet, J. (2025). Comparing the spatial structure of cities in East Africa, Europe and North America. *Habitat International*, *157*, 103329. https://doi.org/https://doi.org/10.1016/j.habitatint.2025.103329

Schiller, G. (2007). Urban infrastructure: challenges for resource efficiency in the building stock. *Building Research & Information*, *35*(4), 399-411.

Simone, A. M. (2004). *For the city yet to come: Changing African life in four cities*. Duke University Press.

Singh, A., Singh, S. K., Meraj, G., Kanga, S., Farooq, M., Kranjčić, N., Đurin, B., & Sudhanshu. (2022). Designing Geographic Information System Based Property Tax Assessment in India. *Smart Cities*, *5*(1), 364-381.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., & Quinn, J. (2021). Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*.

Smith, D., & Crooks, A. (2010). From buildings to cities: techniques for the multi-scale analysis of urban form and function. *Centre for Advanced Spatial Analysis, University College London*.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE*, *10*(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

UN-GGIM. (2018). 14 Global Fundamental Geospatial Data Themes International Workshop on Global Fundamental Geospatial Data Themes for Africa, Addis Ababa, Ethiopia.

United Nations. (2014). Report of the open working group of the general assembly on sustainable development goals. *New York: United Nations*.

Watson, V. (2009). 'The planned city sweeps the poor away…': Urban planning and 21st century urbanisation. *Progress in planning*, *72*(3), 151-193.

Wekesa, B. W., Steyn, G. S., & Otieno, F. A. O. (2011). A review of physical and socio-economic characteristics and intervention approaches of informal settlements. *Habitat International*, *35*(2), 238-245. https://doi.org/https://doi.org/10.1016/j.habitatint.2010.09.006

Yiannakou, A., Eppas, D., & Zeka, D. (2017). Spatial Interactions between the Settlement Network, Natural Landscape and Zones of Economic Activities: A Case Study in a Greek Region. *Sustainability*, *9*(10).

Zhou, Q., Zhang, Y., Chang, K., & Brovelli, M. A. (2022). Assessing OSM building completeness for almost 13,000 cities globally. *International Journal of Digital Earth*, *15*(1), 2400-2421. https://doi.org/10.1080/17538947.2022.2159550

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryClassifyingurbanareasintosettlementclasses.docx